CLASSIFICATION WITH CLASS-INDEPENDENT QUALITY INFORMATION FOR BIOMETRIC VERIFICATION

THÈSE N^{O} 3953 (2007)

PRÉSENTÉE LE 16 NOVEMBRE 2007

À LA FACULTÉ DES SCIENCES ET TECHNIQUES DE L'INGÉNIEUR

INSTITUT DE TRAITEMENT DES SIGNAUX

SECTION DE GÉNIE ÉLECTRIQUE ET ÉLECTRONIQUE

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

POUR L'OBTENTION DU GRADE DE DOCTEUR ÈS SCIENCES

PAR

Krzysztof KRYSZCZUK

M. Sc. in Psychology, Rensselaer Polytechnic Institute, Troy, New York, Etats-Unis
et de nationalité polonaise

acceptée sur proposition du jury:

Prof. J. R. Mosig, président du jury
Dr A. Drygajlo, directeur de thèse
Prof. J. Kittler, rapporteur
Prof. F. Roli, rapporteur
Prof. J.-Ph. Thiran, rapporteur

ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

Lausanne, EPFL
2008

Acknowledgements

This dissertation would not have been possible without the guidance and support of my advisor, Dr. Andrzej Drygajlo. He has been the best boss I could have wished for. My thanks go out to Prof. Fabio Roli, Prof. Josef Kittler and Prof. Jean-Philippe Thiran for their thorough and critical reviews, which greatly helped give this work its final shape.

Research results contained in this work have been greatly influenced by my close collaboration with Jonas Richiardi, with whom I have shared the laboratory space during our entire stay at the EPFL as doctoral students. I also would like to express my gratitude to Anil Alexander, Plamen Prodanov and the rest of my friends and colleagues at the Signal Processing Institute. Working with you over the last few years has been most rewarding.

I wish to thank Prof. Javier Garcia-Ortega and Dr. Julian Fierrez-Aguillar (Universidad Autónoma de Madrid) for generously making parts of the Biosec database available for my experiments during my three-week research visit, realized within the BioSecure Network of Excellence, in 2006. Results of these experiments make an important part of this dissertation. It is also a due moment to gratefully acknowledge the help and support of Dr. Conrad Sanderson and Dr. Sebastien Marcel, who kindly shared with me their knowledge and experience.

Last but not least, my very special thanks go to my parents, Grazyna Lopuszanska and Michal Kryszczuk, for being always there for me, and to Cynthia Delory - apples and oranges made it all click!

The work presented in this thesis was partially funded by grants within the framework of the Swiss National Science Foundation (SNSF) and National Centre of Competence in Research (NCCR) Interactive Multimodal Information Management IM2.

Contents

Resumé

Les systèmes de vérification biométrique d'identité rencontrent fréquemment des difficultés engendrées par des conditions d'acquisition de données non contrôlées. Dans de telles conditions les signaux biométriques peuvent subir une dégradation qualitative due à des facteurs externes et indépendant de l'identité vérifiée. Il a été démontré dans de nombreux rapports qu'une dégradation de la qualité du signal biométrique est une cause fréquente de détérioration de la performance de classification. Cet effet se produit également pour les systèmes de classificateurs multiples et de systèmes multimodaux, lesquels sont systématiquement plus performant que leurs équivalents mono-classificateurs. En cherchant à améliorer la robustesse des classificateurs face à des données dégradées, les chercheurs ont commencé à introduire des mesures de qualité de signal dans les méthodes de classification.

Dans les approches existantes, le rôle de l'information de qualité indépendante des classes est régi par des notions intuitives plutôt que mathématiques, résultant dans une distinction claire entre les méthodes mono-classificateur, multi-classificateur ou multimodales. Les chercheurs ont porté plus d'attention sur l'application de mesures de qualité dans les systèmes de classificateur multiples, avec l'idée sous-jacente qu'un classificateur ayant à disposition des données de haute qualité doit être privilégié par rapport à un classificateur qui opère dans un environnement bruyant. Dans le cas de systèmes mono-classificateurs, la qualité a été utilisée pour la sélection de modèles, classificateurs ou seuils. Dans les deux cas, les mesures de qualité ont la fonction de méta-information qui supervise mais n'intervient pas directement avec le ou les classificateurs utilisés pour classifier les attributs spécifiques a chaque modalité et les attributs sélectifs pour chaque classe.

Dans cette thèse nous argumentons que le même mécanisme régit l'utilisation des mesures de qualité dans les systèmes de classification simple et multiple, et nous présentons une perspective quantitative plutôt qu'intuitive sur le rôle des mesures de qualité dans la classification. Nous notons le fait que pour un ensemble donné d'attributs avec leurs distributions marginales fixées, la séparation des classes dans l'espace joint des attributs change avec les dépendances statistiques observées entre chaque attribut. Le même effet s'applique si cet espace contient des attributs indépendants des classes. En conséquence, nous démontrons que la séparation de classes peut être améliorée en augmentant l'espace des attributs avec des informations de qualité, indépendantes des classes, sous réserve qu'il y ait des dépendances statistiques avec les attributs sélectifs. Nous discutons comment construire des ensembles de mesures de qualité de classificateur dans lesquels les dépendances entre les scores de classification et attributs de qualité permettent de diminuer davantage les erreurs de classification par rapport à un système n'utilisant que scores de classification.

Nous proposons un nouveau cadre théorique, $Q - stack$, qui améliore la classification avec des mesures de qualité indépendantes des classes, basées sur le concept de superposition de classificateurs (*stacking*). $Q - stack$ considère un ensemble de classificateur dans lequel la première couche de classificateur est constituée de classificateurs unimodaux, et la seconde de classificateurs qui opèrent

sur des attributs composés de scores normalisés de similarité et de mesures appropriées de qualité. Nous présentons $Q-stack$ comme un cadre général de classification basée sur l'information de qualité et nous soutenons que les méthodes précédemment proposées de classification avec des mesures de qualité sont des cas particuliers de $Q-stack$.

Nous discutons ensuite du problème de l'estimation de la probabilité d'une erreur unique de classification. Nous proposons d'utiliser l'interprétation bayésienne de la probabilité d'événement unique pour évaluer l'exactitude des décisions de classification. Nous proposons d'appliquer la méthode de prévision des erreurs basée sur la confiance (*credence*) comme une extension du cadre proposé par $Q-Stack$ utilisant un classificateur bayésien superposé. La méthode d'estimation de confiance proposée et la prévision d'erreur héritent de l'avantage de l'incorporation directe d'information de qualité dans l'estimation de confiance. Nous proposons un ensemble de critères objectifs d'évaluation pour les estimations de confiance. Nous discutons comment la méthode proposée peut être appliquée en association a une stratégie appropriée de réparation pour ramener des erreurs de classification à un niveau de cible désiré.

En conclusion, nous montrons une application de $Q-stack$ et de son extension fonctionnelle à la prévision d'erreur liée à la vérification biométrique d'identité en utilisant des modalités de visage et d'empreinte digitale, ainsi que leurs combinaisons multimodales, sur une base de données biométrique réelle. Nous prouvons que l'utilisation des méthodes de classification et de prévision d'erreur proposées dans cette thèse permet une réduction systématique des taux d'erreur à un niveau inférieur à ceux des classificateurs individuels initiaux.

Mots-clefs: reconnaissance des formes, vérification biométrique d'identité, mesures de qualité, prévision des erreurs

Abstract

Biometric identity verification systems frequently face the challenges of non-controlled conditions of data acquisition. Under such conditions biometric signals may suffer from quality degradation due to extraneous, identity-independent factors. It has been demonstrated in numerous reports that a degradation of biometric signal quality is a frequent cause of significant deterioration of classification performance, also in multiple-classifier, multimodal systems, which systematically outperform their single-classifier counterparts. Seeking to improve the robustness of classifiers to degraded data quality, researchers started to introduce measures of signal quality into the classification process.

In the existing approaches, the role of class-independent quality information is governed by intuitive rather than mathematical notions, resulting in a clearly drawn distinction between the single-, multiple-classifier and multimodal approaches. The application of quality measures in a multiple-classifier system has received far more attention, with a dominant intuitive notion that a classifier that has data of higher quality at its disposal ought to be more credible than a classifier that operates on noisy signals. In the case of single-classifier systems a quality-based selection of models, classifiers or thresholds has been proposed. In both cases, quality measures have the function of meta-information which supervises but not intervenes with the actual classifier or classifiers employed to assign class labels to modality-specific and class-selective features.

In this thesis we argue that in fact the very same mechanism governs the use of quality measures in single- and multi-classifier systems alike, and we present a quantitative rather than intuitive perspective on the role of quality measures in classification. We notice the fact that for a given set of classification features and their fixed marginal distributions, the class separation in the joint feature space changes with the statistical dependencies observed between the individual features. The same effect applies to a feature space in which some of the features are class-independent. Consequently, we demonstrate that the class separation can be improved by augmenting the feature space with class-independent quality information, provided that it sports statistical dependencies on the class-selective features. We discuss how to construct classifier-quality measure ensembles in which the dependence between classification scores and the quality features helps decrease classification errors below those obtained using the classification scores alone.

We propose $Q - stack$, a novel theoretical framework of improving classification with class-independent quality measures based on the concept of classifier stacking. In the scheme of $Q-stack$ a classifier ensemble is used in which the first classifier layer is made of the baseline unimodal classifiers, and the second, stacked classifier operates on features composed of the normalized similarity scores and the relevant quality measures. We present $Q-stack$ as a generalized framework of classification with quality information and we argue that previously proposed methods of classification with quality measures are its special cases.

Further in this thesis we address the problem of estimating probability of single classification errors. We propose to employ the subjective Bayesian interpretation of single event probability as

credence in the correctness of single classification decisions. We propose to apply the credence-based error predictor as a functional extension of the proposed $Q - stack$ framework, where a Bayesian stacked classifier is employed. As such, the proposed method of credence estimation and error prediction inherits the benefit of seamless incorporation of quality information in the process of credence estimation. We propose a set of objective evaluation criteria for credence estimates, and we discuss how the proposed method can be applied together with an appropriate repair strategy to reduce classification errors to a desired target level.

Finally, we demonstrate the application of $Q - stack$ and its functional extension to single error prediction on the task of biometric identity verification using face and fingerprint modalities, and their multimodal combinations, using a real biometric database. We show that the use of the classification and error prediction methods proposed in this thesis allows for a systematic reduction of the error rates below those of the baseline classifiers.

Keywords: pattern recognition, biometric identity verification, quality measures, error prediction

List of Figures

List of Tables

Glossary

Terminology and definitions

In this section we provide a glossary of terms used throughout this thesis. Certain terms presented here have a specific use in the area of biometrics. Others are used inconsistently in the topic literature and therefore require a precise definition before being employed in the text of this thesis.

Noise: Any property of the sensed pattern which is not due to the true underlying model but instead to randomness in the world of the sensors [42].

Signal: Observed or recorded physical manifestation of a phenomenon of interest.

Classification feature, feature vector: Extracted from the recorded signal, class-selective properties of the signal. In this thesis we deal exclusively with numerical features. Feature extraction can be seen as a lossy compression of recorded signals, where the compression's objective is maintaining maximal class selectiveness at minimal compression rate. An ordered sequence of features form a feature vector.

Classification score, similarity score, score: Yield of classifier's discriminant function. Usually a descriptor of the position of the observed feature vector in respect to the decision boundary in the feature space.

Classifier decision: Discretized classifier output, predicted class label.

Confidence, confidence measure: a measure of the distance of the test pattern from the decision surface. Confidence measure encodes the certainty of the classifier in making the decision, and is not class-selective. For instance, absolute value of the classification score is a *confidence measure*.

Quality measure: A measure of the impact of noise on the observed signals. Quality measures can originate from the signal or classification feature domains, and are in general class-independent.

Credence: subjective Bayesian probability, or *degree of belief* in an occurrence of a probabilistic event [71].

Acceptance/Rejection: In biometric identity verification, the terms acceptance and rejection refer to *accepting* or *rejecting* an identity claim, and therefore are equivalent to alternative dichotomization decisions. On the other hand, the term *rejection* has been extensively used in pattern recognition literature to describe an action of discarding a particular decision or decisions from the classification process, and therefore refraining from taking a classification decision. In order to avoid confusion in the nomenclature, in this thesis we choose to maintain

the traditional use of the term *rejection* , and we apply it when discussing error rejection techniques, and never to refer to *rejection of an identity claim*. The classes are consistently referred to as *class A* or *class B*.

Probabilistic notation

A large portion of mathematical notation in this thesis refers to probabilistic concepts. Alternative notations for probabilistic events, their instances, etc., exist in the literature. In this thesis we adhere to notation used in [42]. Namely:

Random variables: Random variables are denoted with capital Latin letters, for instance N, X, QM.

Instances of random variables: Instances of random variables are denoted with small Latin letters corresponding to the capital letters that denote the respective random variable. For example, qm denotes an instance of the random variable QM.

Probability and likelihood: Probabilities are denoted with capital letter P, for example $P(x)$ denotes probability of random variable X taking the value of an instance x. Analogical notation for likelihood is $p(x)$.

Probability density functions and distributions: In this thesis there is no functional difference between the use of discrete distributions and continuous probability density functions of random variables. Both probability density functions and distributions are denoted as $p(x)$, where x is an instance of a random variable X. Throughout this thesis, the simplified notation of $p(x)$ replaces the proper notation of $p(X = x)$ for the reasons of notational simplicity, following conventions of standard textbooks in pattern recognition [42, 152].

Conditional probability Conditional probabilities are denoted as $P(Y|Z)$ for *probability of Y given Z*.

Symbols and abbreviations

class A, *class B*, A, B : statistical processes that define alternative classes in dichotomization. In probabilistic notation A and B also denote random variables. For instance, $P(A|x)$ denotes a conditional probability of the instance x being generated by a probabilistic process A, defined by its distribution $p(x|A)$.

C probabilistic event of correct classification decision.

F probabilistic event of incorrect classification decision.

s : observed (recorded signal).

f, $\mathbf{f}$: classification feature, feature vector.

X, x, $\mathbf{x}$: classification similarity score, yield of classifier's discriminant function, instance of a classification score, instance of a vector of classification scores.

τ, $\tau(x)$: decision threshold, decision threshold operating in the domain of x.

Ψ : decision surface in the evidence space.

QM, qm, **qm** : quality measure, instance of quality measure, instance of a vector of quality measures.

N, n : noise, instance of noise.

E, e, **e** : evidence, instance of evidence, instance of an evidence vector.

D : classification decision as a binary random variable.

AC : classification accuracy.

ER, ER_A, ER_B : classification error, error for class A, error for class B.

R : credence, credence estimate.

$\overline{R}$: mean credence.

GT : groundtruth, binary information about the true class alignment of observations.

$\mathcal{N}(\sigma^2, \mu)$: Gaussian function with variance σ^2 and mean μ.

BN : Bayesian Network

DCT : Discrete Cosine Transform

EM : Expectation-Maximization

GMM : Gaussian Mixture Model

$HTER$: Half Total Error Rate

LDA : Linear Discriminant Analysis

MAP : Maximum A-Posteriori (adaptation)

ML : Maximum Likelihood (parameter estimation)

MLP : Multi-Layer Perceptron)

PCA : Principal Component Analysis

QDA : Quadratic Discriminant Analysis

SVM : Support Vector Machine

1 Introduction

*Nogg's Postulate: Any system that depends on reliability is unreliable.**

The beginning of the 21st century was rich in events that turned the world's attention to public security. Progressing globalization gave the human kind possibilities of information transfer and ease of physical mobility unseen before. With those possibilities comes risk of fraud, theft of personal data, or even theft of identity. One of the ways to prevent this is biometric identity verification.

Biometric identity verification is witnessing a stage of a rapid transition from academic and industrial laboratories to real world applications. This transition creates new constraints and new challenges for the emerging biometric technologies, in particular considering their robustness to varying data quality. This demand motivated the work presented in this thesis: a development of a consistent methodology of pattern classification in the presence of degraded signal quality. The findings contained in this thesis are therefore developed directly with the biometric application in mind, however they can easily be extended to any other domain of pattern classification where signal degradation may affect the classification performance.

The problem of biometric identity verification can be portrayed as an example of a two-class classification problem, and as such we treat it throughout this thesis. We introduce the problem of quality of biometric signals and its impact on the accuracy and classification performance of biometric identity verification systems. In this thesis biometric identity verification system serves as an instance of a classifier system that works in a little-constrained environment, where the quality of recorded signals can be compromised. Consequently, we develop a theoretical framework of improving classification accuracy using class-independent information about the quality of classified signals. Acknowledging the fact that the use of quality measures in classification can reduce, but not necessarily eliminate misclassifications, we identify the problem of predicting and rejecting potentially erroneous classification decisions. In order to address this problem we propose a method of estimating the probability of single classifier errors, which can be used together with an appropriate repair strategy for further improvements in classification accuracy.

* popular aphorism in computer science, of unknown authorship.

1.1 What is biometric authentication?

Biometrics is a discipline of pattern recognition where individuals are recognized by their biological or behavioral characteristics. The *identity* of a person is encoded in different biometric traits, routinely referred to as *biometric modalities* [133, 134]. Frequently used types of biometric modalities include: fingerprint, palmprint, face, voice, signature, iris, retinal blood vessel patterns, palm geometry, palm blood vessel patterns, gait, eye movements, etc. [104, 134].

Biometric recognition can be divided into two tasks:

- Biometric identity verification (authentication)
- Biometric identification

Biometric *identity verification*, also referred to as *authentication*, is a task of verifying the authenticity of an identity claim, in other words establishing if a given individual is indeed who he claims to be. Biometric identification is a task of assigning a correct identity label to one person out of a larger pool of individuals.

Biometric identity verification is currently the most popular form of biometric classification. Its applications include identity documents, access control, etc. In this thesis we are concerned with biometric identity verification. We develop a general theory applicable to classification with class-independent quality information, while maintaining biometric identity verification as the head motivation and focal point of the research presented.

Biometric identity verification finds its application everywhere where there is a need for a certainty that an individual is who he claims to be. Such certainty is indispensable in every situation in which an individual claims his rights to certain benefits. Such benefits may include access to financial assets, privileged or private information or services.

Examples of practical applications of biometrics include identity verification for:

- physical and logical access control
- secure identity documents
- forensic applications

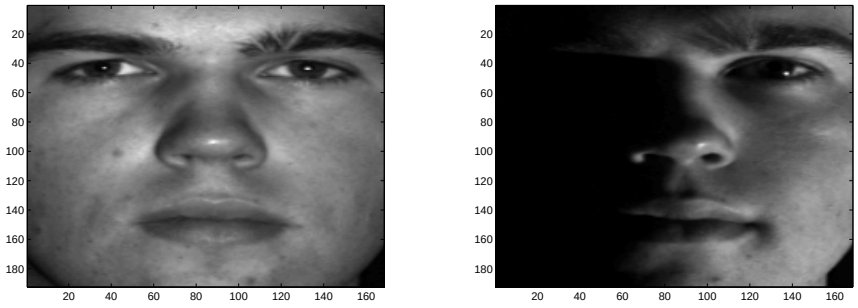
1.2 Unconstrained environments - a challenge for biometric classification systems

A common feature of biometric identity verification applications is that they all involve collecting biometric data from individuals whose identity is under scrutiny at the place and time of the verification. For example, in the access control scenario the identity must be verified at the time of the access attempt. Due to the variability of the environment and behavioral factors the quality of the biometric data collected for the purpose of identity verification may not always be well-controlled and can vary significantly.

Consider two typical examples of environmental and behavioral conditions that may affect the quality of collected biometric data:

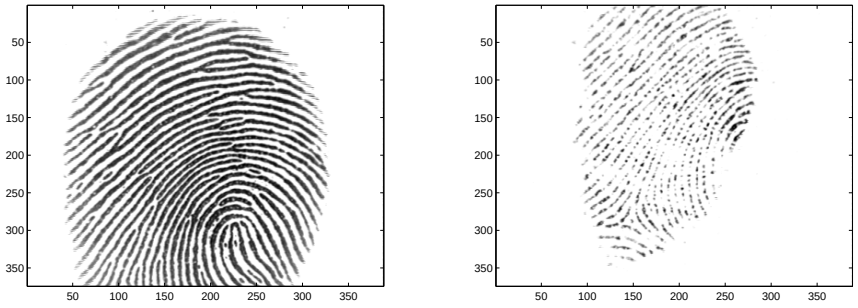
- The appearance of a face in a visible-light image may change dramatically depending on the ambient illumination conditions, facial expressions, facial hair, makeup, etc. [99].
- Ambient temperature and humidity may affect the quality of a fingerprint image collected using an optical sensor [105].

It is easy to understand that a degradation of the quality of a biometric signal may render it less useful or reliable for the purpose of identity verification. Just to follow the two examples mentioned, illumination can change the appearance of a face beyond recognition, and an impression of an excessively dry finger can easily make detection of characteristic fingerprint ridge features impossible. Examples of biometric data samples of high and low quality are shown in Figures 1.1 and 1.2.



(a) (b)

Figure 1.1: Impact of (a) frontal and (b) non-frontal illumination on the appearance of a face. Both images present a face of the same individual and are taken at the same head pose and expression, during one recording session. Example taken from the Extended YaleB database [13].



(a) (b)

Figure 1.2: Examples of fingerprints of (a) high, and (b) low quality. Both fingerprints originate from the same finger. Example taken from the FCV2002-1a database [103].

It has been demonstrated that certain amount of variability in data quality can be accounted for by using different methods of normalization applied to signals, classification features, or similarity scores [11, 77]. Invariably however, if the signal variations are large due to their inconsistent quality, there is a limit to the success of methods that do not explicitly take the quality into account.

The impact of adverse environmental conditions on the characteristics of the collected biometric data can be quantified. For example, the impact of strong directional illumination on the face appearance can be expressed in terms of a ratio between the area of directly illuminated face parts and the area of visible face parts that are covered in shadow. Impoverished quality of a dry fingerprint impression can be expressed by the overall image contrast, etc. In general, the impact of the

environmental conditions may be quantified using *quality measures* [134].

1.3 Measuring quality of biometric signals

Human beings have an intuitive notion of quality of their percepts. Most of us would probably consider a noisy speech signal as low quality. Similarly, a well-defined, sharp and contrasting face image could be well regarded as being of a high quality. This intuitive notion is often in a close relationship with the human internal pattern recognition processing [59, 116]. A noisy speech is hard to understand and hence it is thought of as being of low quality. A blurred image creates problems when trying to recognize a face in it. Many of these naturally-driven intuitive notions of quality will hold when dealing with automatic pattern recognition systems. But not all of them will hold.

Automatic pattern classification systems typically use numerical features that can be considered as compressed versions of the original signal. It is a specific type of lossy compression in the sense that its objective is to preserve as much information as possible about the individuality of the donor of the biometric signal. The compression is obtained by discarding the information that is irrelevant to his individuality. Many of the intuitive quality metrics that humans apply to their percepts would relate to the quality of all perceived information, rather than only to the information that is relevant to the pattern recognition task. A quality measure that is useful for an automatic system would therefore relate to the degradation of the class-selective information, rather to the perceptual qualities of the signal.

1.4 Biometric identity verification with quality measures: a problem or a variety of problems?

Realizing the necessity of estimating data quality in the process of biometric identity verification, researchers turned their attention to quality measures and their role in the classification process. Following two important observations have been made [48, 134]:

- Performance of a classifier may change as the signal quality changes.
- Different classifiers operating on different feature sets are sensitive to different aspects of signal quality.

As a consequence of these two observations, the research work in the field has drifted in two apparently orthogonal directions. The first observation resulted in a number of reported attempts to modify certain parameters of the deployed classifier according to the observed signal quality, thus creating a set of alternative models or classifiers, each of them designed to work well with one very particular range of observed signal quality [49]. Quality-conscious systems that operate according to this paradigm strive to select the closest-to-optimal available classification setup depending on the measured quality of the recorded signal.

The second observation inspired authors to replace a single classifier with classifier ensembles, seeking robustness in their divided competence range. An ultimate example of the latter approach is the combined use of classifiers that operate on different modalities, with the notion that reduced quality of one modality shall not affect the quality of another. The multimodal approach has been baptized as *multimodal fusion* and is currently one of the areas in biometric research that attracts growing attention [48, 78]. In classifier ensembles the quality measures usually play the role of an arbiter - the weight of the opinion of a classifier in the decision of the ensemble is regarded higher if the classifier has high quality data at its disposal.

Looking through the current advances in the area of biometric identity verification with quality measures one could easily have an impression that the two approaches discussed above are indeed different and that the methods proposed incompatible. It therefore appears purposeful and reasonable to take a step back and find a common denominator that governs the use of quality measures in pattern classification. The search of such a common denominator with a particular application to biometric identity verification in mind was the motivation behind the work presented in this thesis.

1.5 Objectives of the thesis

The objectives of the work presented in this thesis are following:

- To propose a quantitative rather than intuitive perspective on the role of quality measures in pattern classification. In this dissertation we are answering the questions why and under what conditions the use of quality information can help to distinguish between classes and consequently to reduce the classification errors.
- To define a general framework of classification with class-independent quality information which encompasses and explains approaches existing up to date in the subject literature.
- To formulate guidelines for developing signal quality measures, which allow for reducing classification error rates in adverse signal acquisition conditions and hence enhance the applicability and robustness of pattern recognition systems, with particular focus on biometric identity verification.
- To develop a method of single-decision error prediction and rejection, compatible with the developed framework of classification with quality information, which would afford further elimination of misclassifications at the cost of employing repair strategies.
- To demonstrate the performance of the developed framework of classification, error prediction and rejection using a real-life multimodal biometric data set containing data of compromised quality.

1.6 From pattern recognition theory to biometric identity verification

The theoretical considerations and developments contained in this thesis are general in nature in order to maintain their relevance to a wide array of pattern classification applications where the problem of signal quality degradation may occur. At the same time however, the application of biometric identity verification is used as a leading motivation in this work and it emerges repeatedly throughout the chapters of this work. Considering our focused attention on the practical applicability of presented findings to the field of biometrics, the evaluation of presenting developments using real biometric data is important. Therefore here a few words of explanation are due regarding the choice of biometric modalities considered.

In this thesis, we chose to consistently focus our attention on the face and fingerprint modalities. The rationales behind this choice are following:

- Face recognition is one of the most natural ways for humans to recognize fellow humans, with dedicated units residing in the brain [59]. It is not surprising therefore that appearance-based face verification systems are attracting high level of interest. This interest is also due to the

fact that face modality is in most cultures the most widely-accepted, non-invasive method of identity verification.

- Fingerprint verification is the first systematically and scientifically researched branch of biometrics [105]. Despite its negative connotations in the criminal context, due to its historical roots and high dependability, fingerprint remains the most popular modality for identity verification.

- Multimodal combinations of face and fingerprint verification are the configurations of choice for currently projected and executed large-scale installations of identity verification systems, including such key applications as travel identity documents [2].

1.7 Main contributions

Major contributions of this thesis include:

- Analysis of the role of quality measures in the classification process and introduction of the concept of classifier-quality measure ensemble. In this thesis we make theoretical advancements towards a better understanding of the role of statistical dependencies between classification features on class separation. We show that given a set of class-selective features, additional class-independent, individually irrelevant features, which sport statistical dependencies on the class-selective features, can improve the separation between classes. As a result, reduced classification rates in comparison to systems that do not use class-independent quality information are theoretically expected, and systematically observed using synthetic and real data sets.

- Development of $Q - stack$, a framework of improving classification using class-independent quality information, based on the concept of classifier stacking. In the developed framework, classification scores originating from a baseline classifier or classifiers are combined with quality measures into evidence vectors and re-classified by a second-layer, stacked classifier. The improved classification performance over the baseline classifiers is due to the existence of statistical dependencies between the class-selective baseline classifier scores and the relevant quality measures. The proposed method provides a unified generalization to existing methods of classification with quality measures and encompasses existing methods relevant to single-, multiple-classifier and multimodal architectures.

- Concept of creating classifier-quality measure ensembles instead of quality measures in detachment from the classification algorithms. We postulate that in order to ensure that a quality measure can be effectively used for classification error reduction it must sport a dependence on the classifier similarity scores. This dependency can be achieved if the quality measure quantifies these qualitative aspects of the classified signals which are also used by the classifier. As an exemplification of this concept, in this thesis we develop two classifier-quality measure ensembles for face modality, and two ensembles for fingerprint modality.

- Development of new quality assessment algorithms for face images. One of the proposed quality measures is a signal-level quality measure that uses an average face template to assess the quality of a face image relative to the quality of a gallery of training images. Another proposed measure is probabilistic in nature and operates on the feature models used by the classifier. This link between the quality measure and the classifier ensures a dependence between the quality measure and the classification scores.

- Development of a probabilistic framework of predicting individual classifier errors taking into account the signal quality information. The proposed method is a natural extension of the proposed $Q - stack$ framework, in which a Bayesian classifier is deployed as the stacked classifier. Individual classification credence is estimated based on the posterior probabilities produced by the stacked classifier. Resulting credence estimates can be used to reject potential misclassifications from the decision pool. For all rejected decisions a repair strategy can be applied, leading to a reduction in classification error.

- Application of the proposed methods of classification and error prediction to single-, multiple-classifier, and multimodal face and fingerprint matching. The proposed method of $Q - stack$ is shown to systematically reduce the classification error rates in respect to the baseline systems. Further, we demonstrate that the single error prediction method presented in this dissertation allows for accurate rejection of potential misclassifications from the decision pool, leading to improved classification accuracy of the remaining decisions.

1.8 Organization of the thesis

This dissertation is organized as follows.

In Chapter 2 we present selected basic concepts of pattern recognition which lay a foundation for further theoretical developments presented in this thesis. We further discuss biometric identity verification as an example of a pattern recognition problem. We provide a detailed discussion of the state of the art in biometric classification with quality measures, with particular attention paid to the face and fingerprint modalities. Finally, we discuss the existing methods of error prediction and classification, with a focus on methods applied in the area of biometric classification.

In Chapter 3 we develop a theory of classification with class-independent quality information. We notice that class-independent quality information can be used to improve class separation thanks to dependencies between classification features, and we provide analytical proofs that this is indeed the case for bivariate normal distributions. Acknowledging the fact that for arbitrary distributions analytical solutions may not be possible to derive, we present $Q - stack$, a statistical, data-driven framework to classification with class-independent quality information. We present the advantages and limitations of the proposed framework using synthetic data sets, and we argue that it provides a generalization of existing approaches towards the use of quality measures in biometrics.

In Chapter 4 the problem of single error prediction and rejection is tackled. We discuss the problem of estimating belief in correctness of a single classification decision, or *decision credence*, from the viewpoint of subjective Bayesian interpretation of probability. As a consequence, we propose a natural extension of the developed framework of $Q - stack$, in which the stacked Bayesian classifier returns credence estimates for classification decisions originating from arbitrary baseline classifiers or classifier ensembles, with or without an involvement of quality measures. We discuss how the proposed credence-based error-prediction scheme allows for rejecting potentially erroneous classification decisions, hence boosting the classification performance at the cost of the necessity of applying a repair strategy to the rejected decisions. We propose a set of evaluation criteria for credence estimates and we demonstrate the performance of the developed method using synthetic data sets.

Chapter 5 gives an evaluation of the proposed methods of classification and error prediction using real multimodal biometric data sets containing face and fingerprint records. In this chapter we discuss the construction of classifier-quality measure ensembles in such a way that the statistical dependencies between the quality measures and baseline classifier scores afford improved class separation in respect to the baseline system. A set of face quality measures is proposed that is designed

to improve the performance of two different baseline face matchers. We show that the framework of $Q - stack$ proposed in Chapter 3 allows for a systematic reduction of the classification error rates below those obtained using the baseline classifiers. In particular, we show that the systematic improvements are observed for single-, multiple-classifier and multimodal architectures of a biometric identity verification system. Further, for the face and fingerprint modalities we demonstrate that the proposed error prediction scheme, as an extension of the $Q - stack$ framework, allows for accurate error prediction and rejection. We evaluate the credence estimates used in error prediction by applying the criteria proposed in chapter 4.

Chapter 6 concludes the thesis. In this final chapter we summarize the findings presented in this work and discuss their generalization to other data types and modalities. We underline the application potential and the limitations of the presented methods and algorithms. Finally, we discuss the prospects for future developments and advancements of the presented methods.

2 State of the art

2.1 Introduction

Biometric identity verification is a pattern classification process of distinguishing between two classes: class of genuine identity claims and class of identity claims originating from imposters [134]. This dichotomization is possible thanks to class-selective features, which originate from a variety of biometric signals that carry information about the individuality of a person. Commonly used examples of such signals include face image, fingerprint image, iris image, but also less known ones are in use, for instance vascular patterns, hand geometry, or behavioral patterns like signature and gait. Following habitual nomenclature in the field of biometrics, we refer to the sources of such identity-dependent signals as *biometric modalities* [134].

Biometric identity verification systems frequently face the challenges of non-controlled data acquisition conditions. In such conditions biometric signals may suffer from quality degradation due to extraneous, identity-independent factors. It has been demonstrated in numerous reports that a degradation of biometric signal quality is a frequent cause of significant deterioration of classification performance [7, 33, 44, 81, 85, 107]. In this thesis we refer to any such degradation as *noise*. Throughout this thesis the term *noise* is used in a very general way, following [42]: we refer to *noise* as "any property of the sensed pattern which is not due to the true underlying model but instead to randomness in the world of the sensors". In the specific context of biometric signals, we use the term *noise* to address any impact of identity-independent auxiliary factors on observed or acquired biometric data samples which degrades the nominal quality of these samples. So defined, the term *noise* encompasses also the typical interpretations of the term encountered in particular in signal processing [75, 79, 147].

Presence of noise in the observed biometric data often causes a degradation of its quality. This degradation can often be measured, and the quality information about observed biometric signals are known as *quality measures* [48, 65, 89, 90, 121, 128, 153]. Quality information is by default class independent and therefore classification is ineffective using quality measures alone. For instance, measuring the amount of ambient additive noise in a recorded speech signal gives an indication of the quality of the signal but it conveys no information whatsoever as to the identity of the speaker.

This thesis is concerned with the use of class-independent information in pattern classification, with special focus on biometric identity verification. In order to give the reader a background for understanding the developments presented in this work, this section contains a compact overview of the related state of the art in pattern recognition and in biometric identity verification, with particular focus on the use of quality information in the classification process. Since in this thesis we pay particular attention to face- and fingerprint-based biometric identity verification, the quality-related problems relevant to these modalities receive our special consideration.

The purpose of a biometric classification system is to minimize misclassifications. However, in a challenging pattern classification application like biometric identity verification errors frequently cannot be avoided. Since errors in identity verifications may carry serious consequences and legal implications for the individuals whose identity might have been erroneously verified, it is of great importance to be able to treat misclassifications appropriately. In particular it is important to estimate the level of trust one can have in a single classification decision. Therefore in this chapter we discuss the existing methods of classifier error prediction, which creates a basis for the discussion of the credence-based error predictors proposed further in this work.

This chapter starts with a recapitulation of the basic concepts of pattern classification in Section 2.2, where we present a conceptual background for biometric verification systems discussed in Section 2.3. Then, common sources of errors in biometric classification systems are discussed in Section 2.4 and the need to collect quality measures is elicited. Section 2.6 discusses in detail the quality-related problems encountered in biometric classification systems with a special focus on face and fingerprint modalities. Existing quality measures are discussed. Next, existing approaches to the use of quality measures in biometrics are discussed in Section 2.7. Finally, we review the existing methods of classifier error prediction in Section 2.8.

2.2 Overview of pattern classification

2.2.1 Basic concepts

Pattern classification is a discipline concerned with assigning measured patterns into classes [152]. The measured patterns are observations and recordings of physical manifestations of phenomena or objects. The nature of observed patterns depends on the measurement type and on the physical properties of the observed phenomenon or object. Pattern classification hinges on the concept of similarity: a *class* is inhabited by patterns which are similar in some sense, or according to some measure of pattern similarity [42]. Pattern classification is concerned with assignment of observed patterns to classes according to the similarity of these patterns with those previously observed, or with those, whose class alignment one has an objective notion of.

The similarity between observed patterns may be estimated based on a comparison of the actual recorded data, unchanged as it comes from the acquisition device or sensor. However, this is often impractical as the observed signals or patterns may contain only certain characteristics which are useful in distinguishing between classes, next to other class-independent characteristics. Therefore the similarity between patterns is often computed based on a selected set of *classification features* which are *extracted* from the observed signals. The feature extraction process can be interpreted as a lossy signal compression, where class-selectiveness and compactness of representation are the compression objectives *. Extracted features are frequently ordered and represented in the form of *feature vectors* [42, 111].

*In many signal compression applications (e.g. image compression), the criteria for evaluating the compression scheme are compactness of representation, compression speed and perceptual fidelity of the signal reconstruction [164]. This is not the case for classification feature extraction, where fidelity of reconstruction is of no importance.

The discipline of pattern recognition emerged from theoretical statistics and as such it inherits many theoretical foundations from this field. Since assigning a pattern to one of the classes is a decision making process, most pattern recognition texts use Bayesian decision theory as a foundation for pattern classification [42, 111, 152], which we also frequently refer to in this work. The Bayesian decision theory is based on the concept of assigning the *most probable* class label to the observed sample [146, 152].

Bayesian decision theory

Consider a task of assigning patterns represented by feature vectors $\mathbf{f}$ to one of the L classes ω_1, ω_2,...,ω_L. Based on prior knowledge, each of the classes is characterized by its probability density function $p(\mathbf{f}|\omega_1)$, $p(\mathbf{f}|\omega_2)$,..., $p(\mathbf{f}|\omega_L)$. Each of the probability density functions can be thought of as a stochastic process which generates feature vectors. Here, $\mathbf{f}$ can assume continuous or discrete values. For discrete features of $\mathbf{f}$ the probability density functions $p(\mathbf{f}|\omega_1)$, $p(\mathbf{f}|\omega_2)$,..., $p(\mathbf{f}|\omega_L)$ are replaced by corresponding discrete distributions. A newly observed pattern is represented by a *feature vector* $\mathbf{f_0}$. The probability that $\mathbf{f_0}$ belongs to class $\omega_l, l \in \langle 1, 2, ..., L\rangle$ is computed using Bayes formula:

$$P(\omega_l|\mathbf{f_0}) = \frac{P(\omega_l)p(\mathbf{f_0}|\omega_l)}{\sum_{i=1}^{L} P(\omega_i)p(\mathbf{f_0})|\omega_i}, \tag{2.1}$$

where $P(\omega_l)$ is a *prior probability* of having a new observation from class ω_l. The conditional probability $P(\omega_l|\mathbf{f_0})$ is usually called *posterior probability* [152]. For the considered L classes, corresponding *posterior probabilities* $P(\omega_1|\mathbf{f_0})$, $P(\omega_2|\mathbf{f_0})$,...,$P(\omega_L|\mathbf{f_0})$ are computed according to Equation 2.1. The feature vector $\mathbf{f_0}$ is assigned to the class whose corresponding computed posterior probability is largest:

$$\underset{i}{\operatorname{argmax}}\, P(\omega_i|\mathbf{f_0}) \rightarrow decide : \omega_i. \tag{2.2}$$

Equation 2.2 describes the *Bayes decision rule*, which is the principle of a *Bayes classifier* [42]. Making classification decisions according to Equation 2.2 takes into account prior class probabilities $P(\omega_i)$. If the prior class probabilities can be assumed equal, $\forall_{i,j}\ P(\omega_i) = P(\omega_j)$, the decision rule becomes

$$\underset{i}{\operatorname{argmax}}\, p(\mathbf{f_0}|\omega_i) \rightarrow decide : \omega_i, \tag{2.3}$$

which is referred to as *maximum likelihood* (ML) decision rule. A classifier creates a *decision boundary* in the feature space [42]. For a Bayesian classifier, the decision boundary between classes ω_i and ω_j is defined by the instances of feature vector $\mathbf{f}$ for which the alternative classes are equally probable, $P(\omega_i|\mathbf{f}) = P(\omega_j|\mathbf{f})$.

The decision rules given by Equations 2.2 and 2.3 are minimum-classification-error rules [42].

2.2.2 Generative versus discriminative classifiers

Functioning of the Bayes classifier hinges on the knowledge of the statistical processes that govern generation of the observed feature vectors $\mathbf{f}$. If these are known, the Bayes classifier is guaranteed to classify observed data with minimal error rates, also called *Bayes error* [42, 111]. The problem is that these processes are usually not known and have to be estimated from available *training* (or *development*) data set. Using estimates of the class prior probabilities and class-conditional feature distributions, a Bayesian classifier can be constructed and applied to classify unseen data, usually referred to as *testing* (or *evaluation*) data.

There are two main approaches towards estimation of class-conditional feature distributions - parametric and nonparametric. The most frequently used parametric approaches include Maximum-Likelihood (ML) and Bayesian parameter estimation. For exhaustive details on the parametric learning methods the reader is referred to [42]. The non-parametric approaches are basically variations of histogram approximations of the underlying probability density functions [152]. In this thesis we use parametric pdf estimation based on the ML principles.

Parametric estimation of a pdf usually requires some prior assumption regarding the nature of the statistical process whose pdf's parameters are under estimation. One of the simplest assumptions is the assumption of Gaussianity of the underlying process. However, for many statistical processes the Gaussianity assumption is inadequate and can lead to a design of a suboptimal classifier. A popular and powerful family of methods of estimating arbitrary pdf is based on weighted linear combination of more than one base function. These methods are called Mixture Models [111]. If a linear mixture of Gaussian functions is used then the resulting models are referred to as Gaussian Mixture Models (GMM). The parameters of the GMM can be estimated using a version of the ML approach, called Expectation-Maximization algorithm (EM) [39, 42]. The GMM modeling using the EM algorithm is used throughout the experiments of this thesis.

The Bayesian classifier operates on estimates of pdfs or distributions which are explaining the statistical processes that generated the observed features. Therefore the Bayes classifier is frequently referred to as the *generative* approach to classification, where the *generative* models define the decision boundary. Note that for the purpose of classification all that is needed is the actual decision boundary, and not the entire form of the generative models. In an attempt to avoid the seemingly unnecessary step of distribution estimation, a series of alternative classification methods have been developed which aim at approximating the optimal decision boundary directly. The most prominent approaches include discriminant analysis methods [42] and Support Vector Machines (SVM) [42, 152]. These approaches are often referred to as *discriminative* since they focus directly on class discrimination, omitting the nature of the feature-generating processes.

2.2.3 Statistical versus heuristic methods

Both the generative and the discriminative approaches to statistical classification are rooted in the basic notion that although the actual statistical processes that generate observed features are unknown, they can be approximated from the available prior observations (training data). Naturally for this notion to be realizable it is required that

1. both the training and the testing data samples originate from the same, unknown feature-generating distributions, and

2. the amount of available training data is sufficient to estimate the distributions (for the generative approaches), or to approximate the class boundaries in the feature space (for the discriminative approaches).

When these conditions are met, the classifier can *generalize* to testing data unseen in the training phase. In some applications, some or both of these conditions can be violated, or be suspected of being violated. For instance, when a classifier is trained on data recorded using one type of sensor and then applied to testing data recorded using another type of sensor, the first condition is likely not to be met. The second condition may not be met for practical reasons if the training data is hard to come by. Both of these problems are frequently encountered in biometric classification.

One must bear in mind that statistical data analysis can only deliver information that actually exists in the available training data, and nothing beyond that. Therefore statistical classification

methods will not generalize to data generated by different processes than these that generated training samples, unless it happens by coincidence. However, in certain instances there may be additional information available about the testing data, which is not present in the training data set. Examples of such heuristics may range from assumptions about distribution differences for data originating from different sensors, to knowledge about what the resulting shape of the decision boundary should be. Examples of employment of such heuristics can be frequently found in the literature, for instance in [7, 49], and we discuss some of them later in this chapter.

We acknowledge the fact that on certain occasions heuristics is a good way of obtaining practically useful results. However, the major problem with methods where heuristics play a deciding role is that they are hardly generalizable to applications and data sets different from those, which these heuristics have been devised for. Therefore in this thesis we do every effort to maintain a strictly statistical approach to classification. The advantage of doing so is that the results and theoretical findings of this thesis are generalizable to unseen data. For practical needs, the methods proposed in this thesis can be always augmented by application-specific heuristic modifications.

2.2.4 Classifier ensembles

There are multiple advantages coming from combining multiple classifiers into ensembles [95, 122]. Classifiers of different nature, even constructed around the same or similar training data, are likely to assign class labels to newly observed feature vectors by applying different decision boundaries in the feature space, each of them being an approximation of the optimal decision boundary. This effect is often referred to as *classifier diversity*, and described in terms of differences in the yields of individual involved classifiers [96, 108]. Combining these approximations by forming a classifier ensemble is likely to return a closer estimate of the optimal decision boundary [155, 156] in the joint feature space of all combined classifiers, thus affording a better generalization properties than the contributing single classifiers [43, 69, 83, 132].

When multiple modalities pertinent to the same physical phenomenon are available, each of the individual modalities can be assigned a separate classifier which is specifically designed to fit the data type of that particular modality. Since the decisions of each of the uni-modal classifiers are related to the same physical phenomenon, it is purposeful to combine these classifiers into an ensemble, which in this case performs a multi-modal classification.

Classifiers may be combined using a plethora of ensemble paradigms, where the individual classifiers in the ensemble can use the same features, a subset of, or a different set of features as the remaining classifiers. Popular classifier ensemble architectures include [122]:

- Bagging [23],
- Boosting [141],
- Stacked generalization (stacking) [165].

Thus constructed classifier ensembles display certain desirable properties that justify their use in spite of their increased complexity in comparison with single classifiers. Such ensemble systems have been shown to generalize well, to be able to handle large data volumes or excessively small training data volumes [122]. Presenting details of multiple classifier theory is beyond the scope or interest of this thesis. Therefore we here limit the discussion of multiple classifier systems to a description of *classifier stacking* [165], because it is the basic concept behind some of the key methods proposed in this thesis.

Classifier stacking - Wolpert's stacked generalization

In his 1992 paper [165], Wolpert declares that *stacked generalization* applies to any combination of classifiers, where output of a classifier is re-classified by a higher-level classifier. In this paper Wolpert refers to classifiers as *generalizers*, so the name *stacked generalization* means as much as classifier stacking, and in this more modern form we shall refer to this method of building classifier ensembles.

The idea behind classifier stacking is this. A classifier may misclassify an observed feature vector if it falls close to the decision boundary, and this misclassification may be due to an improper approximation of the optimal decision boundary. Assume that a sufficiently large number of observed feature vectors are misclassified in the same way due to the same decision boundary approximation problem. Then, a second layer classifier which has both the true class alignment of the data and the actual output of the first classifier can be trained to avoid such systematic misclassifications. In this sense, a stacked classifier reduces the variance in classification decisions of the first-level, baseline classifiers [155], given a proper sub-sampling of the available training data. In this way a stacked classifier ensemble approximates the optimal decision boundary in the entire feature space better than the individual first-level classifiers.

The stacking architecture can be extended horizontally and vertically: the output of an arbitrary number of first-level classifiers can be used as classification features to train the second-level stacked classifier, and the output of a stacked classifier can be as well further given to a third-layer stacked classifier or classifiers, and so on. Within the stacking architecture, classifiers can be combined using a wide array of rules, from simple binary decision-based voting through weighted linear combinations, to trained score-level schemes [20, 43, 83, 131].

The concept of classifier stacking is widely used throughout this thesis. First, all multiple-classifier ensembles applied to single modalities and multimodal trained fusion ensembles are straightforward instances of classifier stacking. Then, the framework of classification with class-independent quality information proposed in this thesis is also based on classifier stacking, where output of the first-level baseline classifiers, augmented by class-independent quality measures, is then used as classification features for the stacked classifier.

2.2.5 Dichotomization

A special type of classification problems are two-class problems, where an observed sample can be assigned to one of two classes, A or B, but not both. A classifier that is assigned to a two-class classification problem is frequently referred to as a *dichotomizer*. Dichotomization has a privileged position in pattern recognition because it is argued that a multiple-class problem can be represented as a series of two-class problems [151].

Biometric *identity verification* is also a dichotomization problem, since its task is to distinguish between two classes: genuine identity claim and imposter identity claim. In the following section we will focus specifically on biometric identity verification as an instance of a two-class pattern classification problem.

2.3 Biometric identity verification as a pattern classification problem

2.3.1 Biometric identity verification

Biometric identity verification systems have the goal of establishing if a biometric observation collected from an individual who claims certain identity matches the data previously recorded from the individual in question. A diagram of a biometric identity verification is shown in figure 2.1.

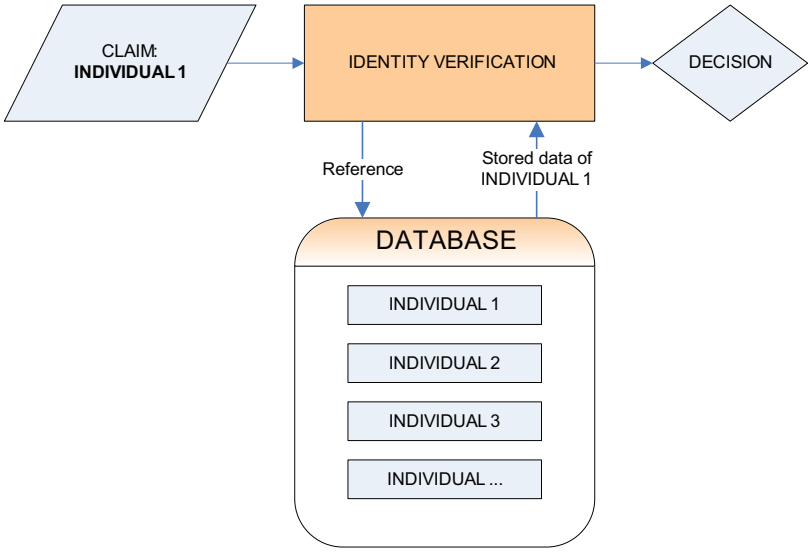


Figure 2.1: Biometric identity verification system. The identity verification module sends a reference identity request to the database and in return receives stored reference data pertinent to the claimed identity. Based on the similarity between the claim and stored reference data a verification decision is taken.

The biometric data collected from the person who claims to be Individual 1 is compared to the data of Individual 1 stored in the database. In order to establish a frame of reference for the comparison, the database usually contains biometric data collected from other individuals. The comparison is performed in the *Identity Verification* module, which determines whether the degree of similarity between the claimant data and the corresponding database information is greater than the level of similarity between the claim data and the data of the rest of the population. Usually, the population is represented by the data of the remaining Individuals, whose data is stored in the database. If the level of similarity between the claimant data and the corresponding data from the database is greater than the similarity between the claimant data and the rest of the population, the decision of the system is affirmative (identity verified, claim *accepted*). Otherwise the decision is negative (identity not verified, claim *rejected*). These two decision possibilities define the two classes in which the biometric identity verification system operates as a pattern recognizer.

Biometric verification systems are therefore examples of two class pattern classification problems (dichotomizers): the claimant either is who he claims he is (accept a genuine claim), or he is not (reject an impostor attempt). In this thesis, we shall consistently refer to these classes as

- **class** A, the class of imposter identity claims, and
- **class** B, the class of genuine identity claims.

In biometric identity verification there is no reason to presume that a claim is more likely to come from an imposter than from a genuine client [137]. Therefore in this thesis we consistently assume equal prior probabilities for classes A and B, $P(A) = P(B)$. Consequently, the optimal Bayesian dichotomizer applied to a task of biometric identity verification operates using the maximal likelihood decision rule given by Equation 2.3.

As opposed to the identity verification task, the problem of biometric *identification* is a multi-class problem. Despite the apparent difference in the problem definition, a multi-class problem of identification can be reduced to a series of two class identity verification problems [115, 151]. In this thesis we are concerned explicitly with identity verification.

2.3.2 Information flow in a biometric system

Biometric identity verification is a process that takes as input acquired biometric signals and assigns to them a class label (*accept* or *reject* the identity claim). Figure 2.2 shows a diagram of data flow in a typical biometric identity verification system, from the stage of data generation until the stage of decision-making.

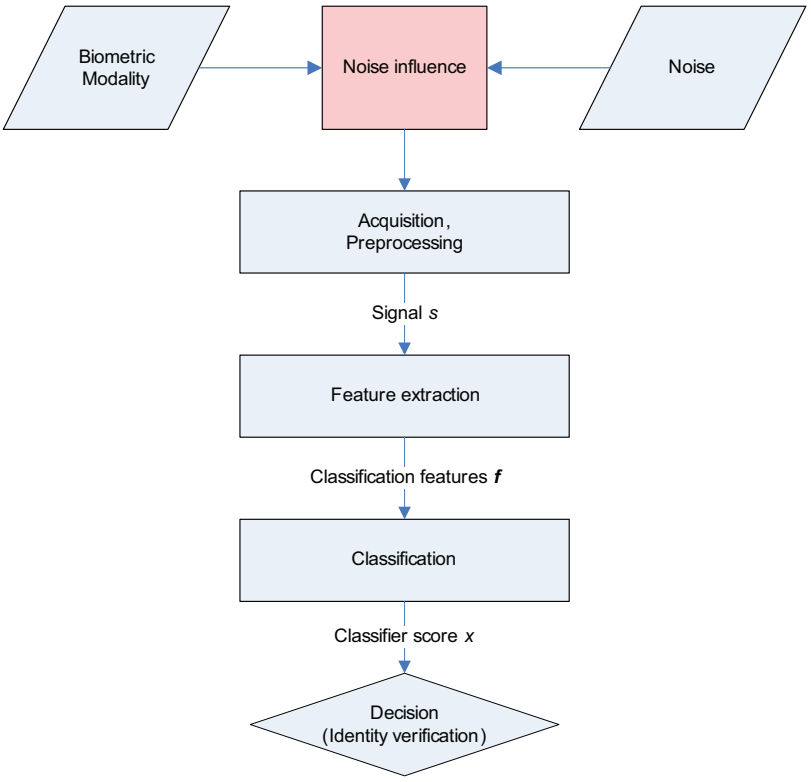


Figure 2.2: Diagram of a typical biometric classification system.

An individual presents to the biometric system his biometric data. Depending on the biometric modality (type of biometric trait used for identity recognition), the biometric data may be affected by extraneous noise of various nature. In the case of speaker verification it can be acoustic noise, in the case of face verification unfavorable illumination conditions, and in the case of fingerprints noise can be of behavioral type - for instance insufficient pressure on the sensor or unstable finger positioning. The noise contamination happens before or during the data acquisition process. Since this thesis is not concerned with the data acquisition process, here we treat the interaction between

the noise and the biometric signals as a black box: a closed process that the system designer has no influence upon but whose parameters can be observed and used in order to improve the classification performance. The data collection process typically involves dedicated biometric sensors - for instance in the case of face recognition it is a camera, in the case of fingerprint recognition it is a fingerprint scanner.

Acquired signals may now undergo modality-specific normalization treatments. For instance face images need to be geometrically and photometrically normalized before further processing [64, 140]. Fingerprint images must be separated from the background and the resulting image is subjected to image enhancement procedures [105, 150]. Thus treated signals are typically passed to the feature extraction process which strips the signal of information that is not relevant to the classification task. The type of normalization and feature extraction techniques depend mostly on the biometric modality, and on the type of classifier employed in the system. Certain biometric systems do not employ any normalization methods, or classify the acquired signals directly, without an explicit feature extraction routines [136]. In such case, the entire acquired and/or normalized signal is equivalent to a feature set or vector. In this thesis we are not concerned with the normalization or feature extraction processes. We assume that these procedures have been performed according to the best practices for the considered modalities. In this work our interest is focused on the pattern classification process.

The actual classification begins here. The extracted features f are passed to a set of routines that compute the yield x of a *discriminant function* Ψ. Following commonly used nomenclature in biometric classification we shall frequently refer to the value of x as *classifier score*, or *similarity score* [42, 152]. In order to arrive at the final classification decision, classification score x is binarized, often by comparing it to a threshold τ, resulting in a binary classification decision. Biometric identity verification frequently involves comparison, or *matching*, of biometric data samples in order to establish if they originate from the same individual. Therefore biometric identity verification system is also referred to as *biometric matcher* [29, 149].

2.4 Classification errors and quality of biometric data

The major difficulty that the biometric systems have to face in real life, out-of-laboratory conditions is the possible degradation of signal quality due to the extraneous noise influences. The term *noise* is here used in a broad term and it encompasses any signal that carries no class-specific information and that intervenes with the actual biometric signal. The noise degradation happens before or during the signal acquisition stage.

The noise contamination masks and changes the selective individual signal features which allow for a successful classification of the biometric signal. If a classifier is trained on clean, noiseless data its performance is likely to suffer from a noise degradation of the signals it is deployed to classify.

We distinguish three types of classification errors in biometric identity verification [128]:

1. *Systematic errors* are those caused by design problems inherent to the pattern recognition system engineering task. These include wrong assumptions about the form or family of the distributions of features under consideration, poor choice of features leading to excessive overlap between classes, insufficient amount of training data, poor estimation of classifier or model parameters, etc.

2. *Presentation-dependent errors* are those caused by unforeseen variability in the signal source. These can be caused by degraded environmental conditions (illumination variation for face, specular reflection for iris, additive noise or channel noise for speech, residual fingerprints

traces, etc.), or by extra variability in a signal (elastic skin distortion for fingerprints, expression of the face, badly executed signature, etc.).

3. *User-dependent errors* happen only with certain users that do not fit the otherwise correct assumptions about the user population. This is a well-known problem in biometrics, and one of its incarnations in speaker recognition tasks is called the Doddington Zoo effect [41].

The systematic errors are not directly linked to the quality of classified signals and as such shall not be a concern of this thesis to any greater extent than that we shall follow the best practices in the field in order to avoid such errors as much as feasible.

The presentation-dependent errors are directly linked to the problem of signal quality and possible deviation of the signal from a nominal quality. Therefore the techniques presented in this thesis will be relevant to the goal of minimizing this type of error.

The task of biometric identity verification has been portrayed as a two-class classification problem. From this viewpoint there is no distinction made between the individual system users. If one wishes to take into account the individual user characteristics then genuine and imposter identity claims pertinent to each user can be considered a separate two-class problem that has no inherent individual-related variability [17]. Therefore without a loss of generality in this thesis we will not focus on the user-dependent errors. For an in-depth analysis of the user-specific issues in biometric identity verification the reader is referred to [48, 117].

2.5 Dealing with low-quality data

The presentation-dependent errors of a biometric classification system are strictly linked to the excessive variability of the observed data. This variability is related to the concept of *signal quality*. In general, when certain characteristics of the observed signals are kept within a tolerable range of variation, such signals are regarded to be of *nominal*, or *high quality*. When the signal characteristics cease to meet the tolerance, the signal is considered to be of *degraded* or *low quality*. The concerned signal characteristics and tolerable range of their variability depend on the considered domain. In pattern classification, the characteristics refer to the class-selective classification features, and the tolerable range is defined by the sensitivity of the classifier deployed to classify these features.

Presentation-dependent misclassifications occur when either the classes are inherently overlapping, or when the quality of the observed signals is degraded. A series of methods has been devised to cope with the problem of excessive class overlap. They include:

Signal normalization: Methods of signal normalization are devised to reduce the signal variability before the classification features are extracted. Usually, these methods take advantage of the fact that certain factors that degrade the signal quality always transform the signal from a high- to low-quality state in a similar way. It is therefore possible to reduce the degrading effect by reversing it. Instances of signal normalization include face image processing [140, 143], or fingerprint binarization [30].

Robust feature extraction: The stability and invariance of classification features is directly linked to the classification performance. In order to ensure that the classifier's performance does not suffer from changes in signal quality, features robust to signal quality variation have been proposed. Instances of such features include DCT features for face recognition [137] and MLP-based robust feature extraction for speaker recognition [139, 166].

Feature and score normalization: Analogically to the process of signal normalization, feature normalization procedures are aimed at reducing the impact of noise on extracted features before

the classification process. After the features have been subjected to the classifier, normalization of classifier scores can be applied in order to minimize the impact of noise before the binary decision is taken [7, 48, 77, 117, 138].

Use of quality measures: If an external factor degraded the observed signal in such a way that classification performance is compromised, then intuitively the notion of the actual signal degradation can be used to compensate for this effect. A series of methods of using quality information to improve classification performance have been proposed. They will be discussed further in this chapter.

All presented methods of dealing with presentation-dependent errors can be deployed to a large extent independently of each other. In this thesis we focus our attention on the topic of the use of class-independent quality measures in classification. However, the methods and algorithms proposed in this thesis do not interfere with the methods of signal, feature or score normalization, and do not prevent a deployment of any robust feature extraction scheme. The use of quality measures is intended as a post-classification step to address these errors that could not be coped with by the pre-classification normalization procedures.

2.6 Quality measures for biometric signals

The concept of biometric signal quality is understood intuitively. For example, when we try to recognize or identify other people by their faces from photographs we would without hesitation say what factors decide about the image's quality: resolution, contrast, sharpness, etc. Similarly, when it comes to latent fingerprint inspection, experts would consider well-formed, rich in detail finger imprints as being of high quality. However, if it is an automatic biometric identity verification system that is expected to use the available signals, the intuitively understood quality metric that humans would apply, may be insufficient or no longer applicable [73]. Although it is definitely some sequence of signal processing, feature extraction and pattern recognition steps that happen in the brain when people recognize other people and other shapes [116], there is no reason to suppose that these steps are replicated by an automatic classification system, especially that there exists a pronounced difference in the face recognition accuracy achieved by humans and by automatic face recognizers [6]. Therefore automatic quality measures must be relevant to automatic pattern recognition.

In general, quality measure of a biometric sample must quantify its suitability for automatic classification and the actual used biometric matcher [65]. The term *suitability* here means that the automatic matcher must be able to extract relevant features from the observed signal and then assign a correct class label. This definition explicitly links the quality measurement to the classification system that uses the biometric signals whose quality is measured.

In this thesis, the term *data quality* is used in a broad sense and it refers to any quantifiable characteristics of the data, but not of its donor. Consequently, a *quality measure* is any quantitative metric that depends in its value on the quality of the data but not explicitly on the identity of the data donor.

To be precise, quality measures may actually sport certain amount of class-dependency. Namely, the way in which the noise interacts with the actual biometric signal may in a subtle way depend on the individuality of the donor of the biometric signals. For instance, an excessively dry finger will produce a low-quality image when presented to most optical fingerprint scanning devices [105], and this degraded quality can be captured in the form of a quality measure [29]. Since some individuals have inherently drier fingers than others, this in itself class-independent quality measure

can appear to have certain dependence on the user identity. Similar links can be found for most quality measures. However, these effects are negligible in comparison with the class selectivity of biometric signals themselves. For this reason, in this thesis we consider information about biometric signal quality to be by default *class-independent*.

In the following sections we review the existing methods of estimating quality of biometric samples for face and fingerprint images.

2.6.1 Face image quality measures

Factors that impact face appearance in an image

Face image may seem to be the most common and familiar of the biometric modalities and its practical applications are numerous. Yet, despite this apparent familiarity, appearance-based face recognition is a particularly challenging and difficult problem in pattern recognition. The appearance of a face in the image will alter as a consequence of a multitude of interactions between the ambient illumination conditions of the image acquisition, behavior of the subject whose face image is recorded, and quality of the acquisition equipment itself [99].

- **The conditions of image acquisition**. A face image is a result of the interaction of light emitted by one or more light sources, and the reflectance of the face surface [64]. While the reflectance of the face is largely characteristic for an individual, the illumination conditions are definitely not. Factors such as number, color, intensity and directionality of light sources will dramatically impact the appearance of the illuminated face. Also, a face is a non-Lambertian surface and for these reasons effects like self-shadowing and specular reflections of the face surface further amplify the influence that the lighting conditions have on face appearance [86].

- **Behavioral factors**. Human face is a non-rigid, deformable object. Even in best-controlled illumination conditions changes in face expression and pose will change the way a face appears in an image. Also, images of persons in motion are significantly harder to acquire than when they do not move.

- **Acquisition hardware and software**. Quality, type and settings of the imaging apparatus are crucial for the output of the face image acquisition process. Resolution, focus, chromatical and spherical aberrations, depth of field, exposure, contrast and color balance, just to name a few, are factors that have a deciding influence on how a face appears in the image. Nowadays, many images are stored in digital, compressed format. Most existing image compression schemes aim at preserving maximal subjective fidelity of an image, rather than preserving all information that is in the original. Therefore there exists a possibility that certain information pertinent to the automatic face recognition process may be lost during compression [98].

Measurement of the impact of these quality-affecting factors, in particular of those that are not specific to face imagery only, have been addressed by the image processing community [79, 147]. However, existing automatic measures of image quality have an objective of modeling human perception of image quality rather than assessing the suitability of the signal for a pattern recognition task [9]. Considering the plethora of factors that may affect the appearance of a face in a captured image, measuring its quality relevant to automatic face recognition is a most challenging task.

Overview of state of the art in face quality measures

Despite the abundance of face recognition systems and the popularity of the subject in the biometric research community, relatively little work has been reported on the problem of measuring quality

of face images. The reason for this situation is that it is extremely hard to pinpoint what aspects of the quality of a face image are of actual importance to the process of face recognition, and up to date there is no consensus about it among the domain experts [27].

Possibly the first documents that explicitly address the problem of face image quality for the purpose of automatic processing is the Technical Report of ICAO "Biometrics deployment of machine readable travel documents", Annex A - Photograph guidelines [2]. The document contains practical suggestions and guidelines regarding collecting digital photographs for travel documents, and it addresses the possible reasons for face image fidelity degradation rather than the topic of image quality itself. For instance, it specifies the minimal size of the image, what percentage of the image should the actual face occupy, it suggests that the image should not be under- or over-exposed, that the subject should not be wearing any headdress, etc. These recommendations as such are rather vaguely defined and by no means guarantee that a face image collected according to these guidelines will be of use for an automatic face recognition algorithm.

To the best of our knowledge the first report on the use of explicit measurement of certain aspects of face image quality in order to improve the performance of an automatic face recognition system was given by Kryszczuk and Drygajlo in [86] and in [85]. In [86] a measure of face area left after an image segmentation procedure was used in order to improve a face matcher based on local DCT features [137]. The quality measure was used in order to adaptively select the optimal, quality-dependent decision threshold. In [85] the signal- and classification feature-level quality measures were introduced in order to avoid erroneous classification decisions when processing noisy face images. The same set of face quality measures was applied to Bayesian networks-based method of decision reliability estimation in [93]. The reliability estimates were applied to perform a multimodal fusion of face and speech decisions.

In the face of the political decisions regarding the introduction of biometrics in the identity documents the problem of face image quality has attracted industrial research. In a paper [73] Hsu et al. used a set of ISO recommendations regarding face image quality [4] as an inspiration for developing a set of automatic quality measures for error prediction in automatic face verification systems. Unfortunately no details of neither the matching algorithm used, nor the implementation of the quality measurement are given. The authors show that there exists a correlation between their quality measures and the classification scores, and that upon a rejection of low-quality samples the classification error rates drop. Hsu et al. noticed justly that quality measures should be capable of predicting classification performance. However, without any available details on the algorithms of neither the classifier nor the quality measures it is impossible to generalize the results from this work.

An intuitive rather than quantitative understanding of the concept of face image quality prevails in [121], where a long list of quality measures is given, albeit again without any implementation or algorithm details. The quality measures include: frontal, rotation, reflection, illumination, spatial resolution, bit per pixel, focus, contrast, brightness, reliability and a combined overall quality measure being a combination of the previously listed measures. Experiments reported by the authors of this work suggest that there is a benefit of using quality measures as a classification feature when performing multi-classifier fusion, but why such improvement would be observed is left without explanation. Unfortunately, without any algorithmic details regarding these rather ambiguously named quality measures it is impossible to analyze their actual role in the reported experiments, or to reproduce them.

In [88] and [87] the idea of predicting classifier errors using quality measures has been further extended. Rather than attempting to predict misclassifications from the quality measures themselves, a method of estimating probability of error based on modeling error-conditional evidence was

presented in [87]. The evidence consisted of signal- and score-based quality measures. As the signal-level following quality measures were used: the normalized Pearson's coefficient of 2D correlation of the query image with an average face template built using a PCA reconstruction, and an image contrast measure. As score-level evidence the distance from the decision threshold and a sum of log-likelihoods of the client- and the world model were used. The quality measures were developed and tested using the face part of BANCA database [10].

2.6.2 Fingerprint quality measures

A fingerprint image is an impression of the ridges and valleys which appear on an individual's fingertip. The overwhelmingly most common impression type are latent fingerprints *, which are the deposits of substances present on the finger surface left upon contact with another (usually solid) object. Latent fingerprints and their use for identification of individuals have been described as early as 1892 by Galton [57], but there exists evidence that the idea itself is far older than that. Currently, fingerprint-based identification of individuals is a well-established discipline in forensic sciences. With the introduction of fingerprint scanning devices the application of fingerprint comparison extended from forensic- and police applications to other domains, like access control and secure documents.

Latent prints are said to be an "inevitable source of error in making comparisons" as they generally "contain less clarity, less content, and less undistorted information than a fingerprint taken under controlled conditions, and much, much less detail compared to the actual patterns of ridges and grooves of a finger" [167]. The same is true for fingerprint images collected directly from the finger using dedicated fingerprint scanners. It is therefore of great importance to be able to quantify the actual quality of the captured fingerprint image before fingerprint comparison (matching) is performed. In the case of latent fingerprints the experience and knowledge of human experts allows them to take into account the fingerprint quality aspects when examining a collected impression. In the case of automatic fingerprint recognition and matching systems dedicated algorithms must do this job.

Quality of fingerprints in the context of pattern recognition

Before we move on to describing the actual algorithms that assess fingerprint quality, a short introduction to the features used in automatic fingerprint matching is due. Obviously, only these aspects of fingerprint quality that are linked to the features used for fingerprint matching can impact the matching performance.

Automatic fingerprint matching can be realized using various features extracted from the fingerprint images. Frequently, experts distinguish between three fingerprint feature levels [105]:

- Level 1: characteristic configurations of global ridge patterns.
- Level 2: location and type of characteristic points in ridge and valley patterns, commonly referred to as the *minutiae*. The most commonly used types of minutiae include *ridge endings* and *bifurcations*.
- Level 3: fine details of the local shapes of ridges. Most well-known level 3 features in fingerprints are the *sweat pores* [76, 92, 148].

The level 1 features are not sufficiently selective to be on their own reliably used for fingerprint comparison and are routinely outperformed by matchers using level 2 and level 3 features [49].

* *latent* (lat.) = "hidden", or "invisible".

Instead, they are routinely used for fingerprint classification and rapid retrieval from large databases [80]. Level 2 features have been historically first to be used for fingerprint recognition and matching, and are currently at the base of the most popular fingerprint matching algorithms [105]. Level 3 features, described in the literature as early as 1912 [102] are a growing area of research [76, 148] and have been shown to augment the discriminative information of level 2 features, but successful localization of level 3 features requires very high resolution and quality imprints [92]. Such image resolution is not offered by most existing fingerprint capture devices, and is not featured by currently available large-scale evaluation databases.

All the features used for automatic fingerprint matching are quite obviously strictly linked to the actual ridge structure. Therefore it does not come as a surprise that quantification of fingerprint quality for the purpose of automatic matching is equivalent to capturing how well the ridge structure of the original fingerprint is replicated in its digital image, and how well can the important elements of the ridge structure, like the ridge orientation and ridge edges, be automatically detected.

Overview of state of the art in fingerprint quality measures

Simon-Zorita et al. [144] list some of the factors that may compromise the quality of a fingerprint impression. These factors include:

- Incomplete fingerprint capture: a significant part of the interest region of fingerprint is out of the capturing area of the sensor.
- Excessive, insufficient or irregular pressure over the sensor device.
- Sensor residues from previous acquisitions.
- Excessively dry, wet, or oily finger.

These factors are causes rather than visible effects of potential degradation of image quality and as such are hard if not impossible to quantify based on the actual fingerprint image alone. Consequently, also the authors of [144] did not propose any automatic algorithm to assess these conditions: instead, they used human observers to give their subjective assessment of the fingerprint image quality.

More promising results are reported by authors who, instead of trying assess the presence and then to guess the possible causes of fingerprint quality degradation, attempted to actually quantify the quality of the fingerprint impression itself. Most of these works are focused on the assessment of local and global coherence of the ridge flow, and on the clarity of ridge separation. Coherent ridge flow and clear ridge separation are necessary for successful detection of fingerprint features, which in turn is a necessary condition for reliable fingerprint matching. A summary and an attempted taxonomy of fingerprint quality measures based on fingerprint image processing is given in [8]. Below we discuss in detail most influential approaches.

Ratha and Bolle [125] propose to detect if a fingerprint image has got clear, well-separated and good-contrast ridges by analyzing the image in wavelet domain. The authors' main motivation is that a popular standard of fingerprint compression, Wavelet Scalar Quantization (WSQ), already demands that the fingerprint be represented in the wavelet domain, and consequently they propose to assess its quality without the need of performing full decompression. The authors suggest also that a non-compressed format might undergo some of the compression steps in order to be compatible with their quality estimator. The main idea behind the presented algorithm of quality assessment is that the Discrete Wavelet Transform (DWT), one of the steps of WSQ compression procedure, encodes various subbands of spatial frequencies. The distribution of energy over different spatial frequency

subbands is different in low- and high quality fingerprints. Namely, in high-quality fingerprints the energy is concentrated in few subbands, while in low-quality fingerprint images this is less the case. The estimate of fingerprint quality is therefore based on computing relative difference of cumulative energy over selected subbands.

LinLin Shen et al. [142] suggest that this may not be a computationally effective approach and propose to assess the fingerprint quality using Gabor filters. One of the characteristic features of Gabor filters is that they are orientation-sensitive, i.e. they produce strong response to directional patterns. Fingerprint ridges are such patterns, and the authors argue that high-quality fingerprint ridges should evoke strong response of these filters, whose inherent orientation agrees with the local ridge orientation. If a fingerprint is of compromised quality and local ridge orientation is weakly pronounced then consequently no directional Gabor filter should produce a strong response.

Lim et al. [101] reiterates Shen's opinion that the approach proposed in [125] may not be suitable in an automatic fingerprint verification system, albeit without giving much reason why not. They relate to the work of Shen [142] in saying that "[...] as the gray level structure of valleys and ridges contains much more information than orientation alone [...] the local directional strength alone may not be sufficient to measure the quality of fingerprint image". In order to alleviate this problem they propose two independent quality measures: a local and a global one. Both quality measurements are based on performing Principal Component Analysis (PCA) on the local gradient vectors in a neighborhood of 32×32 pixels. The dominant eigenvector thus computed describes the local ridge orientation in the chosen neighborhood. In the local quality assessment a periodic variation of pixel intensity along a line perpendicular to the dominant local eigenvector is used. In the global quality assessment the smoothness of the changes of local ridge orientations is considered. The authors propose to combine the local and global measures into one by computing their linear combination.

In the respective evaluations the above discussed fingerprint quality assessment schemes made no link with the task of automated fingerprint recognition, although they all claimed that the development of their quality measures was motivated by such tasks. The necessity of making a link between the quality measures and classification was clearly stated by Tabassi et al. in [150]. In this work, the authors describe an automatic quality assessment algorithm, whose implementation has been included in the publicly available fingerprint matching software package, NFIS2 [162]. The authors begin with an assumption that a quality measure must be a predictor of the similarity score obtained in the process of automatic fingerprint matching process. The authors argue that if for matching high-quality fingerprints the similarity should fall high, and for low-quality fingerprints it should fall low. Since the quality measure, logically, should be high or low as a function of the actual observed fingerprint quality, it should be a predictor of the similarity scores.

The authors take a pattern classification approach towards assessing the fingerprint quality. Namely, they a priori predefined five discrete quality levels, where level 1 corresponds to excellent and level 5 corresponds to poor quality. The particular levels of quality are defined by obtained classification error rates on a preset fingerprint database. The number of the discrete levels was chosen rather arbitrarily. These five preset quality levels become classes to which each newly observed fingerprint is to be assigned to. The feature vector for this classification consists of information on the number of high quality minutiae. The quality of each of the minutiae is determined based on the statistics of pixels in the immediate neighborhood of the minutiae (in order to quantify the local contrast), and on the position of the given minutiae on a fingerprint quality map. This map is computed taking into account factors like fingerprint contrast, ridge flow, and ridge curvature. The minutiae itself are detected using the very same detector algorithm that is used later by the actual matcher that the quality measurement scheme is designed to work with. The authors report an epic journey in search for a proper decision function that would assign the minutiae quality features to

one of the five quality bins. As the result of their lack of success in finding an explicit formulation of such function, they left the job to a multi-layer neural network. Further publications of [65, 149] basically describe the same idea.

There is, however, one serious flaw in the first assumption taken by the authors. Namely, if quality measures are to be predictive of the matching scores then a comparison of two fingerprints of high quality that come from different fingers would have to result in a high matching score, which would lead straight to an obvious misclassification. Similarly, comparing two imprints of the same fingerprint, assessed to be of low-quality, would necessarily produce a low matching score, and again a misclassification. Further, this assumption causes the definition of fingerprint quality somehow circular - fingerprints that have been misclassified are used as training sets to extract quality features that later serve as predictors of quality, and according to the previously set assumptions they must predict if these fingerprints will be misclassified. Nevertheless, making the link between quality measures and classification scores is an important contribution of this work.

In [29] Chen et al. proposed fingerprint quality measures that predict the classification performance rather than the similarity scores themselves, which in contrast to the assumption presented in [65, 149, 150] is a conceptually sound approach. Indeed, it can be intuitively expected that matching low quality fingerprints can result in increased number of misclassifications, without a presumption what kind of errors they are going to be. In this work the authors propose two quality measures, one based on a frequency analysis, and another based on an analysis of local ridge gradients. The quality measure based on the frequency analysis is conceptually related to the method proposed in [125], but instead of computing DWT coefficients the authors propose to apply computationally efficient Fast Fourier Transform (FFT). There exists a clear difference between the spatial power spectra for high-and low quality fingerprints. Namely, a well-defined ring can be found in the spectrum of a high quality fingerprint, which is a consequence of regular and well-define ridge spacing. This ring is not present for low-quality samples, where image distortions make the power spectrum much more uniform. A set of band-pass Butterworth filters is then used to quantify the amount of energy present in each of a pre-defined series of power bands. The entropy of the distribution of the responses of these filters measures the degree of local energy concentration, and as such is used as the quality measure. The second quality measure described in [29] is conceptually similar to that proposed in [101]. The authors propose to compute the eigenvalues of the covariance matrix of a matrix of fingerprint pixel intensity gradients in a local neighborhood. The resulting quality measure is a measure of the relative difference between the first and the second eigenvalue. Large value of the proposed quality measure means that there exists one dominant local ridge orientation in the considered neighborhood. This hints on high quality, well-defined ridge structure. The authors further show that indeed their proposed quality measures are good predictors of classification performance on selected databases.

2.7 Classification with class-independent quality information

By analogy to the traditional general division of classification methods into single classifier and classifier ensembles, here we present the prior art divided into single classifier and multiple classifier systems that us quality information. The bottomline of this division is not only the amount of classifiers involved, but also the general logic between both approaches. The authors who reported on the use of quality information in single classifier systems did so seeking to optimize certain quality-dependent parameters in the classification process that eventually led to lowering of the error rates. The typical role of the quality measures in a multiple classifier system is to help in selecting or appropriately weighting the contributions of each classifier involved in the ensemble:

without actually trying to improve the performance of any of the single classifiers themselves.

The proposed methods can be also conveniently categorized as *heuristic* and *statistical*. In the heuristic methods the notion of the influence of quality degradation is encoded into the classification scheme by its designer. In the statistical approaches, this influence is learned from the available data. It must be noted here that the boundary between the statistical and heuristic methods is not crisp. Namely, it is difficult if not impossible to avoid bringing in certain amount of heuristics into the statistical classification schemes. For example, often the choice of the family or parameters of statistical models is clearly an ad-hoc, heuristic procedure that is based on the prior knowledge of the domain. Since there is no clear criterion. Here we propose to apply a rather conservative view on the distinction between the heuristic versus statistical methods. We refer to all methods that do not involve a training phase as heuristic, the remaining ones being statistical.

In general, given sufficient availability of training data, statistical methods bear the potential to outperform the heuristic approaches [43]. It must be noted however, that there is nothing intrinsically inferior about heuristic classification methods: if the heuristically chosen models or model parameters accurately represent the data, such classifier will be optimal in terms of classification accuracy.

2.7.1 Single-classifier systems with quality measures

The application of quality measures to single-classifier systems is at the first glance not very intuitive. Quality measures by themselves are in general par excellence class-independent. However, it has been noticed that degradation of signal quality does effect the classification error rates. The reason for such degradation in classification performance is that the statistical distributions of class-selective features tend to change as the signal quality changes. Understanding this shift is not intuitive either and probably therefore all reported methods of using quality information in single classifier systems apply some type of heuristics and simplifying assumptions to handle the complexity of this problem.

The impact of varying signal quality on the classification accuracy has been noticed by the forensic speaker recognition community. A discrepancy between the qualitative attributes of different data pools is here referred to as *channel mismatch* or *condition mismatch* [7]. Reynolds [127] demonstrated that training several probabilistic Gaussian Mixture Models (GMM) for different speech channel conditions, and then performing a channel-dependent procedure of model selection helps prevent recognition errors due to the channel mismatch between training and testing data. The most appropriate model is chosen to be the one that returns the highest likelihood for the new observation. The likelihoods can be compared directly since they all are produced by an iterative adaptation of one model. In this way, the quality is not explicitly measured, rather the quality information is encoded and quantized in the process of training individual channel-dependent models.

In a way, the procedure proposed by Reynolds could as well be considered as a multiple classifier system since the model selection process could be interpreted as a choice between a set of alternative classifiers. However since no information originating from two models concurrently is used, we regard it as a single classifier method.

In [7] Alexander proposed a conceptually related method of performing distribution scaling in order to account for a channel mismatch in speaker verification. Namely, the author suggested to estimate the scaling parameters by comparing the distributions of similarity scores obtained for data of different quality. The scaling transformation is then applied to the scores computed for the testing observations. Again, the quality measures are not computed explicitly, the quality of the data is considered constant within the separate data corpora used to compute the statistical score transformation parameters.

Campbell et al. [25] proposed to use information that can be considered as quality measures: the speech-to-noise ratio, channel type and the duration of a speech utterance, in order to estimate the confidence of a speaker verification decision.

In [85] a segmentation scheme for non-frontally illuminated face images was presented. The scheme allows for feature extraction from only the part of the face image that was not severely affected by the non-frontal illumination. The authors demonstrated that a quality-dependent adaptive thresholding strategy allowed for reducing classification errors on a face verification task.

In a report regarding the fingerprint recognition in the framework of the US Visit program Wein and Baveja noticed that modifying the decision threshold according to the signal quality allowed for decreasing the classification error rates [163]. Here, the authors based their findings on the results obtained using the publicly available NFIS2 fingerprint verification software [58], which includes the fingerprint quality assessment functionality NFIQ [150]. The same quality measurements were used by Grother and Tabassi [65] to predict performance of the fingerprint matcher. The authors tried to find a function that maps the quality estimates of a matched pair of fingerprints onto the score space. The authors recognize that the existence of a dependence between classification scores and quality measures is prerequisite for quality-based error prediction but the consequences of this dependence are neither properly analyzed nor exploited for increased classification accuracy. The authors propose only an error prediction and rejection scenario for low quality of classified samples. Similar approach is taken by Hsu et al. in [73] where a set of rather vaguely described face quality measures is shown to be predictive of classification scores. Unfortunately no details regarding the classifier used are given, which renders the reported results impossible to reproduce or analyze in greater detail.

2.7.2 Multiple-classifier systems with quality measures

An early introduction of quality measures to the process of combining classifier yields was presented by Chatzis et al. [28]. The authors focus on devising two methods of quality-based fuzzy clustering for combining outputs from three face-based and one speech-based classifier. The title of this paper may be misleading: although it announces decision-level fusion, in fact the fuzzy clustering techniques do not use binary decisions but rather continuous classification scores ("soft" decisions). Out of the two proposed techniques, the *Fuzzy Data k-Means* (FDKM) and the *Fuzzy Data Vector Quantization* (FDVQ) only FDKM shows some error reduction over the respective counterparts without quality measures. This improvement, due to the introduction of quality measures, is present but not for every combination of the used classifiers. Unfortunately, neither the nature of the quality measures used, nor their distributions are discussed in the paper. Additionally, the authors merely state the results but do not provide any discussion on why such results were obtained with respect to the used data. In this situation it is difficult to learn a lesson or generalize from presented results. However, the results do hint on the fact that including quality measures in the process of combining classifiers into ensembles does have its merit and deserves a closer investigation.

Toh et al. proposed a method of incorporating quality measures in the multiple-classifier score fusion based on Multivariate Polynomial Regression [153]. The authors accurately argue that linear weighting methods may ignore possible nonlinear dependencies in the data and evaluate a second- and a sixth-degree polynomial score combination schemes. The quality measures are used to modify corresponding weighting terms in the polynomials. Presented results are rather inconclusive and hardly support the authors' claim that the introduction of quality measures improved the performance of the classifier ensemble. The authors mention in the paper that quality measures may constitute an additional degrees of freedom which ought to help to separate the data. As we will show further in this thesis, indeed that is the case.

In [46], [51], [50], [21] and [48], Fierrez-Aguilar et al. presented two score-level methods of incorporating quality measures into the process of classification using multiple classifiers. The authors use the term *adapted schemes* to describe the proposed algorithms, in the sense that certain parameters of the function that combines the scores originating from separate classifiers change according to a quality measure. Namely, the authors propose two approaches, a *generative* and a *discriminative* one.

The generative approach is based on prior work by Bigun et al. [18, 19]. The authors assume that the class-conditional scores from all classifiers are normally distributed. Consequently, the authors construct two *score functions*, which in fact appear to be the parametric estimates of the class-conditional score distributions. Then, the expected values of the class-labels as predicted by both class-specific score functions are computed. Finally, the combined score is found by selecting greater of the two expected values. The application of the quality measures in the process is in the estimation of the variances of the scores produced by single classifiers. In [19] the classifier variances are estimated from the training data. Instead, in [48] Fierrez-Aguilar proposes to substitute the variance estimate with an inverse of the quality measure q_i: $\upsilon_i^2 = 1/q_i^2$. The quality measure is defined as a single scalar, namely a geometric mean of the quality measures derived from the two compared signals.

There are a few problems with this approach. First, while certain experimental data may well support the assumption of the Gaussianity of the class-dependent score distributions [19] this needs not, and is generally not the case. The proposed method in its current form does not offer any possibilities of adjustments for other than normal score distributions. Second, the proposed method of substituting the score variance by an inverse squared quality measure lacks mathematical foundations. However, this heuristic procedure may have certain intuitive appeal to it: higher quality data suggests a small score variance, therefore assigning more importance to that particular score in the fusion process. Third, the method requires that one and only one quality measure is used in the multi-classifier fusion process. Also, the normalization of quality measures is asymmetric: the quality measures between the *lowest* and the *normal* (presumably nominal) are mapped to a fixed interval from 0 to 1, but the *highest* value of the quality measure seems to be allowed arbitrarily large values greater than 1. It must be mentioned that the term *generative fusion method* is a misnomer here: the quality measures are not at all used in the training process, in fact they are not modeled at all and their use is heuristic in nature.

The discriminative approach to multi-classifier combination is proposed in [46, 48, 51]. The proposed method is developed based on the principles of Support Vector Machines (SVM) [42]. The quality measures come to play when training the classifier by optimizing a quadratic term given by Equation 3.14 from [48]. In order to introduce the quality measures a constant weight term C_i from Equation 3.14 is replaced by a term that is proportional to the product of quality measures pertinent to all involved classifiers (here equivalent to modalities since intra-modal fusion is not explicitly accounted for). The intuitive rationale behind this approach is to give more importance to the support vectors that have higher associated quality rating. The implementation, however, forces that this weighting be linear - a change in quality measure will impact the weights in a proportional way. Note that the quality measures do not impact the vector $\|\mathbf{w}\|$. Only one quality measure per classifier/modality is considered. In the presence of multiple quality measures one could imagine including them all in the product term in Equation 3.41 in [48], but then this would be equivalent to considering different quality measures equally important and giving them equal leverage on the weighting. There is no reason for such assumption and in fact in this thesis we provide arguments against it.

Another type of heuristics was employed by Fierrez-Aguilar et al. in order to combine scores

originating from two fingerprint matchers [49]. Here the authors noticed that one of the fingerprint matching algorithms is more robust to the fingerprint image degradation than the second one. Following this observation the authors propose a heuristic method of quality-weighted score summation. Namely, with the decreasing quality of the observed fingerprint image subject to matching, the fusion scheme relies more weight on the system that is more robust to the quality degradation. Again, the solution has its intuitive explanation but there is no good reason to think that the proposed scheme is even nearly optimal. As authors themselves state in the conclusions, the method cannot be easily generalized to other data sets, classifier combinations and quality measures. Generalization to multiple quality measures is also not obvious.

We return to the topic of classification with quality measures in Chapter 3, where we present a general theoretical framework that encompasses the use of quality measures in single-, multiple-classifier and multimodal architectures.

2.8 Predicting and handling classifier errors

2.8.1 Basic concepts behind error prediction in pattern classification

The task of a pattern recognition system is to correctly assign observations to classes. Good practice in classifier design suggests ways to construct classifiers in such a way that the number of erroneous classification decisions is minimized. Errors in classification may be as much due to inadequate classifier structure or parameters, but also to the inherent structure of the classification problem (e.g. overlapping classes). The problem of designing an optimal classifier that minimizes the number of misclassifications has been extensively studied. In particular, methodology of estimating the error bounds is well known [42].

Theoretical estimation of the error bounds roots in the assumption that the statistical properties of the processes that generate observations are known. In practice that is often not the case and only approximations of the underlying statistical processes can be found. In order to minimize the effects of approximation error, a variety of empirical methods of estimating expected classification error has been devised.

Vast majority of those methods demonstrate how to arrive at an estimate of the expected average error rate of the classifier. In other words, the estimate gives an approximation of how many times the classifier is expected to commit an error over a large volume of observations. In other words, every observation is assigned the same, average probability of being correct.

Intuitively in most situations certain classification decisions can be expected to be more certain than others. The basic reason for the variation of the certainty of the decision lies with the fact that in most applications of pattern recognition the observations are usually generated by stochastic rather than deterministic processes. The importance of the exact notion of the certainty of each classification decision is very much application and context-specific. For example, in a large-scale mass production quality control application, the accuracy of each individual quality control check may not be as important as the overall error estimate.

Prediction of an average classification error gives an overview of the expected performance of the system over a sufficiently large pool of testing data. In many applications such a notion is sufficient since success or failure of each individual classification attempt is not very important, as long as the total error rate remains within the estimated error bounds. An example of such an application may be quality inspection of products in a manufacturing plant.

In certain applications, however, accuracy of each single classification decision is of high importance, and a misclassification may carry severe consequences. Examples of such applications may

be rare event detection (aerospace, cataclysm, finance). Another example of such an application is biometric identity verification, where a decision concerns a particular person and therefore its certainty is of particular interest and importance. In such applications, one is interested in not only the average error rate of the system, but also in predicting misclassifications for each single decision. We therefore seek a system that delivers an estimate of the probability of the correctness of a single decision, while taking into consideration the quality of the classified signals.

2.8.2 Predicting classification errors

If the statistical class-conditional distributions of the observations are known, it is possible to derive corresponding error bounds for discrimination between the given classes. It is guaranteed that the lowest achievable error rate, hence also the average probability of error, can be obtained using the Bayes decision rule [42, 111]. The error bound approximation is given by the Chernoff and (looser) Bhattacharya bounds. Thus derived error approximations rely strictly on the knowledge of underlying statistical distributions and priors, and have no room for auxiliary information (quality measures).

As early as 1957 C. K. Chow discussed the concept of discarding (rejecting) potentially erroneous classification decisions, in the context of automatic character recognition [31]. The work was strictly empirical and without theoretical foundations. Such foundations were presented later, when Chow introduced the concept of Error-Rejection Tradeoff in pattern classification [32]. According to this work, before actually being assigned a class label, an observation may be accepted, or rejected from the process of classification. For a given system, its performance in terms of the error and rejection rates is claimed to be completely defined from the statistical distributions of the observations. The problem of computing the probability of a single decision error is not explicitly addressed, and one can only infer from presented derivations that on the decision level, observation-conditional posterior probabilities would be used to accept or discard an observation from actual classification. Also for a two-class problem, like biometric identity verification, the method is of questionable application since the reject rate is always zero for any decision whose posterior probability is above the chance level (0.5). There is no room for observing classifier's behavior and no possibility of making use of class-nonspecific information.

The idea of discarding potentially inaccurate decisions has been adopted by Cordella et al. but in their work the concept of reliability remains rather intuitive than probabilistic, and without any account for a potential condition mismatch estimator [34, 35]. In [62] Golfarelli et al. build upon Chow's concept of the Error-Rejection Tradeoff, with the application to biometric verification systems (hand shape and face). Here, the authors again use the class-conditional posterior probabilities as the estimates of the degree of belief in the decision correctness, or as dichotomizer's confidence measure. However, here the decision threshold used in the decision rule is not applied to withdraw an observation from the classification process due to insufficient confidence. Instead, the rejection of the sample is equivalent with assigning one of the classes (identity not verified). Using thus reformulated decision rule the authors derive the error rate approximations that characterize the dichotomizer. The average probability of decision correctness is computed, rather than the probability of arriving at an accurate estimate of the probability of the correctness of single decisions. There's no account for auxiliary information that may help to discard reliable from unreliable decisions. The subject of rejecting single classification decisions is further treated in [52, 54, 55]. In all of the reports mentioned here that deal with error rejection, only total classification accuracy is considered, and not class accuracies (or, equivalently, class errors). This simplifies the problem, but in certain applications of pattern recognition, like biometrics, such simplification may lead to too-far going conclusions, and errors need to be considered on a per-class basis.

Above approaches are based on the assumption that the statistical distributions that generate observations subjected to classification are known, or that can be approximated with a satisfactory accuracy. In this sense, the predicted error rates are theoretically derived, without falling back to observe how the classifier actually performs. In contrast, another group of techniques takes a more pragmatic stand: in order to predict future performance of a classification system, the error rates on a dedicated data set are observed. Such a paradigm allows for a direct estimation of the probability of error, in accordance with the classical probability definitions. Namely, the probability is computed based on the actual observed number of correct and incorrect classifications.

A survey of methods applied to estimate and predict error rates in pattern classification can be found in [70] and [154]. In general, such methods devise data sets for classifier construction and training (training set), and for error estimation (development set). The development set is frequently dissected in multiple ways for the purpose of more reliable estimation of the expected error performance. Typical strategies include resubstitution, train-and-test, cross-validation (with leave-one-out strategy as the special case) and bootstrapping (selection of subsets with allowed data repetition) [154]. Mentioned methods invariably focus on the estimation of systems' error rates, and no attention is paid to the error probabilities of single classification decisions.

Described methods have been also frequently applied to error prediction of biometric systems. Wang and Bhanu propose to learn the distributions of fingerprint match and non-match, observed on a training data set [160]. A parametric model is trained using the EM algorithm and applied in order to arrive at the system's expected upper and lower error bounds. The same authors presented a model of error rate prediction for fingerprint- and ear shape-based biometrics, based on error integrals obtained from observing matching scores for biometric samples of varying quality [161].

An intuitive way for estimating classifier's performance is based on analyzing the margins - the absolute difference between the dichotomizer's accuracy in choosing one class over another, observed on a development data set [120]. Margin-derived confidence measure is a function of the distance from the decision surface, whose yields are contained in the $\langle 0, 1\rangle$ range. However, the probabilistic meaning of the differentially-defined margins is not clear, and therefore margins, as such, do not provide any measure of the actual probability of the event of a single classification being correct. Also, no explicit account for class-nonspecific information can be taken.

In his doctoral dissertation Poh proposes to predict the performance of biometric verification systems, in terms of expected error rates and DET curves [117]. Probability of making single correct decisions is not addressed.

Error rates, or average error probability estimates give a prediction of the expected performance of a classification system on a large testing data set. Each single decision does not receive particular attention and is therefore assigned the same, average probability of error (or being correct). There are, however, good reasons to try to identify single classification decisions that are more likely to be correct from those that are likely to be incorrect. These reasons include:

1. In some applications, single classification decisions are important and therefore knowing the certainty at which the decision is arrived at is beneficial. Examples may include biometric identity verification (e.g. forensic scenarios), financial decisions, rare event prediction (natural disasters) [45].

2. Decisions that are identified to be potentially incorrect may be further dealt with on individual basis. Decision repair strategies may include decision rectification, repeated signal acquisition and classification/model adjustment [124, 129].

Only a handful of authors attempted to handle the problem of identifying potentially erroneous classification decisions.

An attempt to predict misclassifications using Post-Recognition Analysis Technique (PRAT) with application to face recognition was described by Li et al. in [100]. At the base of the proposed method, the ratios of "Failure Prediction False Alarm Rate (FPFAR)" and the "Failure Prediction Miss-Detection Rate (FPMDR)" are defined. However, the ratios are difficult to understand or interpret in a probabilistic sense. Namely, the FPFAR is defined as a ratio of the total number of observed false rejections over the sum of the numbers of the observed false rejections and correct accepts. Similarly, the FPMDR is defined as a ratio of the total number of observed false accepts over the sum of the numbers of the observed false accepts and correct rejects. Once defined, these ratios seem not to be used at all throughout the paper. Instead, authors propose to use AdaBoost to classify features derived from the sorted similarity scores. Resulting prediction is therefore strictly binary and can not provide any insights into the probability of single accurate or inaccurate classification decision. It should be also noted that boosting techniques are known to require rather large volume of training data, and that is often not the case in biometric applications. The authors do not take into account the possible use of sources of information other than classification scores for error prediction.

Most recently, Peng et al. proposed to use the concept of perfect recognition to analyze the classifier's performance on the task of face recognition [113]. The essence of their approach is to use the very same data set twice - once to train the classifier, and then to test it. The authors claim that the similarity scores obtained on the test using training data encode completely the structure of the classifier. Furthermore, according to the authors it is always possible to achieve 100% accuracy on the training data set: in most cases, particularly in biometrics, this claim seems to be difficult to realize without heavy overfitting of the classification models. The authors suggest that the thus derived "Perfect Recognition Similarity Scores" (PRSS) allow for a prediction of classification performance without the use of additional query (development or testing) data. Namely, the non-matching (biometric claim rejection) scores are assumed to follow a single normal distribution. Each score is assigned a normalized distance from the mean of this distribution. Finally, an exponent of thus computed distance measure, divided by an arbitrarily chosen parameter λ , called metric f, is averaged over the entire set of training scores, yielding a metric of classifier's performance. An average system error rate is then predicted by fitting Gaussian distributions into sets of PRSS and ARSS (Actual Recognition Similarity Scores), where the later ones are derived using a development data set, thus contradicting the initial claim that such data is inutile for performance prediction. In order to predict the success or failure of individual classification decisions the authors employ a Support Vector Machine (SVM) classifier, whose features are derived from the f metrics. The sign of the output of the SVM classifier predicts the correctness or incorrectness of the single observation in question, therefore one still cannot infer the probability of the prediction. Also, there is no means of incorporating extraneous, class-nonspecific information that may help to predict recognition failures.

Indeed, auxiliary class-independent information can prove to be predictive of the classification errors. Grother and Tabassi [65] declare that "[biometric] sample quality [is] a scalar quantity that is related monotonically to the performance of biometric matchers [...]" and suggest that with an accuracy determined by a residual error ϵ_{ii}^k a function of quality P should approximate the actual matching scores. Further in the paper, the authors admit that simple regression techniques for estimating P proved to be inaccurate when applied to fingerprint matching, and related $NFIQ$ quality measures [58, 150]. In fact, function P is a way of expressing the dependency of matching scores on the quality measures, and there is really no reason to make the aforementioned assumption of monotonicity, as we have shown in Chapter 3. The authors however need this assumption in order to be able to find a monotonic error-rejection tradeoff for their case [32]. The resulting error-rejection tradeoff curves show a direct dependence between recognition errors and the particular

quality measures they used, which in turn has no probabilistic meaning. For this reason, as the authors themselves honestly admit, the applicability of their method to other quality measures, and consequently also other matchers or modalities, is unknown.

An effort to address the issue of probabilistic error prediction has been recently undertaken at the Swiss Federal Institute of Technology Lausanne (EPFL). In a series of publications two parallel approaches to address the problem have been presented: based on graphical models proposed by Richiardi et al. [130], and based on explicit Gaussian Mixture Modeling proposed by Kryszczuk and Drygajlo [87]. In [87, 130] the measure of probability of a single decision correctness is referred to as *reliability measure*. Here, the use of the term *reliability* is strictly functional since it relates to the dependability of the obtained classification decisions, or their ability to perform the desired operation in the system. From the mathematical perspective, thus formulated problem of decision reliability estimation is equivalent to the estimation of *degree of belief*, or *credence* in a single error correctness. This topic will be treated separately further in this chapter.

Both approaches suggest modeling of error-conditional distributions of evidence, where evidence contains both class-specific information and quality measurements. In [130] Richiardi et al. propose a Bayesian network (Figure 2.3) to model the dependencies between the evidence and the classification scores used for credence estimation:

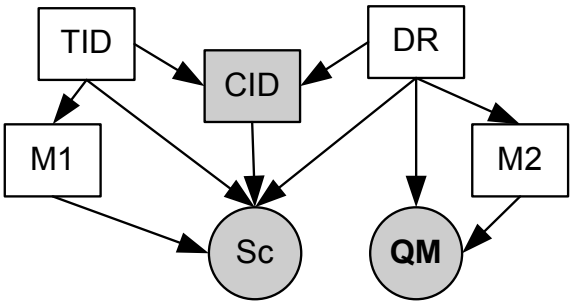


Figure 2.3: Bayesian network used by Richiardi et al. to model evidence for credence estimation.

As Figure 2.3 shows, the evidence model is split into the quality measure model $M2$ and the score model $M1$. Consequently, the credence values are computed as follows:

$$\begin{aligned} P(DR=1|cid,sc,\mathbf{qm}) =& \alpha P(DR=1)P(TID=cid) \\ & \cdot \sum_{M1} P(M1|TID=cid)\cdot P(sc|DR=1,M1) \\ & \cdot \sum_{M2} P(M2|DR=1)\cdot P(\mathbf{qm}|DR=1,M2). \end{aligned} \tag{2.4}$$

From Equation 2.4 and from the structure of the graphical model shown in Figure 2.3 it is clear that the classification score-level evidence is considered independent of the quality measures. This assumption, however, is difficult to justify, and in the GMM-based approach proposed by Kryszczuk et al. the dependencies between evidence are implicitly modeled by a full-covariance matrix of the error-conditional multi-gaussian model. In [130] Richiardi et al. consider the reliability estimates as a classification confidence measure and use them in order to improve the performance of a single classifier system by the means of a sequential repair strategy. A similar method, also using BN, has been used by Kryszczuk et al. to tackle the problem of handling erroneous unimodal classifier decisions and consequently fuse them [93, 94]. The main problem with the approaches adopted

in [87, 93, 94, 130] is that the reliability estimates were computed using error-conditional, rather than class-conditional evidence distributions. While the use of error-conditional distributions has sound theoretical explanation and an intuitive appeal, modeling of such distributions, in particular using the normality assumption [130], is usually more difficult and less accurate than modeling of class-conditional distributions. This topic will be handled in more detail in chapter 4 of this thesis.

2.8.3 Probability of error as a single event probability

Classical methods of predicting classification performance return a single probability value as the error predictor. This value may be arrived at either by theoretical means (Bayes error bound and similar) or empirically. In the later case, regardless of the implementation details or cross-validation method used, essentially the error rate is found by comparing the volume of erroneously classified observations α to the total observation volume under classification β:

$$P(Decision = Error) = \frac{\alpha}{\beta} \tag{2.5}$$

This equation computes probability in the classical sense. Note that this estimate is rather a *descriptor* than a *predictor* - it observes the true state of nature (actual class alignment of the observation) *a posteriori*, after the observations have been classified. Upon encountering a new observation, all that can be done is to assume the same, constant error probability $\frac{\alpha}{\beta}$.

How to really *predict* the outcomes of a single, newly observed data sample? We need to find an estimate of the probability of a *single event* rather than assuming that all future events are equally probable. According to the classical interpretations of probability, however, a probability of single event is not defined. According to Equation 2.5 if all we have is one observation, then its probability of being correctly classified can be only zero or one - definitely not an answer we search for. Let us look for a different than classical interpretation of probability that would let us estimate the probability of a single classification decision.

2.8.4 Interpretations of probability

The concept of probability dates back to Christiaan Huygens who gave the first known scientific treatment of the subject [74]. In general, the concept of probability refers to a measurement of chance of a particular event. Most popular interpretations of probability include [71]:

- **Classical Interpretation:** Laplace assigns probabilities in the absence of any evidence, or in the presence of symmetrically balanced evidence. The guiding idea is that in such circumstances, probability is shared equally among all the possible outcomes, so that the classical probability of an event is simply the fraction of the total number of possibilities in which the event occurs. The classical definition of probability requires an absolute knowledge of the domain, which is not realistic in the pattern classification domain where class samples rather than entire population data are available. Further, the classical interpretation of probability does not include the concept of single event probability.

- **Logical Interpretation:** Probabilities can be determined a priori by an examination of the space of possibilities. However, they generalize it in two important ways: the possibilities may be assigned unequal weights, and probabilities can be computed whatever the evidence may be, symmetrically balanced or not [26]. Not applicable to the problems treated in this thesis since the exploration of the entire space of possible classification outcomes is not feasible.

- **Frequency Interpretation:** Classical interpretation counted all the possible outcomes of a given experiment, finite frequentism counts the *actual* outcomes. Frequentists view probability of an event as its relative frequency over time, i.e. its relative frequency of occurrence after repeating a process a large number of times under similar conditions [159]. This is also known as aleatory probability. Not applicable, since it does not encompass the concept of probability of a single event.
- **Propensity Interpretation:** Probability is thought of as a physical propensity, or disposition, or tendency of a given type of physical situation to yield an outcome of a certain kind, or to yield a long run relative frequency of such an outcome [123]. The concepts of propensity are also used to predict an outcome of a single event [60, 61]. Not applicable in this thesis since the propensities of a signal that are related to its potential to be correctly classified without any posterior knowledge are hard if at all possible to identify.
- **Subjective probability (subjective Bayesianism):** Treats probability as a *degree of belief*. We identify probabilities with *degrees of confidence*, or *credences*, or *partial beliefs* of suitable agents, who must act rationally [135]. The subjective Bayesian interpretation of probability is compatible with the notion of single-event probability, and therefore it is used throughout the rest of this thesis.

The subjective Bayesian interpretation of probability therefore allows us to compute the single event probabilities as degrees of belief that a single classification decisions are correct. The reasonable estimation of such degree of belief can be done knowing both the state of the nature in the past (which can be considered equivalent with the definition of probability in the classical sense) and the pieces of information about the current observation that may hint on how likely we are to not go wrong when classifying it. We shall refer to this information about the currently classified sample as *evidence* and denote it as E. According to the Bayesian interpretation, we seek the probability of making a correct classification decision, given the evidence E. We return to this subject in Chapter 4, where we propose a systematic approach towards estimating single decision credence.

2.9 Summary

In this chapter we have introduced biometric identity verification as a pattern recognition problem. We have discussed the steps leading from an acquisition of a biometric signal to the final classification decision, and we have mentioned the impact of extraneous noise on each of the steps of the classification process. We have introduced the concept of biometric signal quality and quality measures. We have discussed in detail the problems connected with deriving quality measures for biometric signals, with a particular focus on the face and fingerprint modalities.

We have discussed the existing methods and algorithms of using quality measures in biometrics. We have shown that these methods treat quality information from an intuitive rather than quantitative perspective. Consequently, we have discussed a rather artificial division into single-classifier-related and multi-classifier-related methods, which is clearly seen in the topic literature. As a consequence of the intuitive understanding of the information that quality measures carry, the discussed state-of-the-art methods are mostly heuristic in nature. In chapter 3 we present a novel view on the problem of classification with quality measures which bridges the gap between these two seemingly incompatible families of heuristic classification methods, and intuitive heuristics will be substituted with a data-driven approach.

In this chapter we have also discussed existing methods of predicting classification errors. We have identified existing methods which allow for an approximation of average system error, and

we have explained why such methods are of limited interest for biometric identity verification. We have presented existing approaches towards single error prediction methods and we have shown that they do not allow for a straightforward inclusion of quality information. Hence we introduced the need for a probabilistic error prediction method that allows for estimating the probability of single classification error, while taking into account signal quality measures. We have discussed existing interpretations of probability and we have shown that the Bayesian subjective credence, or degree of belief, is an interpretation that best suits the set objectives. Consequently, a credence-based method of error prediction is proposed in Chapter 4

Classification with class-independent quality information

3

3.1 Introduction

In Chapter 2 we have discussed the basic concepts in dichotomization and we have introduced the concept of quality measures. We mentioned how a change in the qualitative properties of recorded signals frequently affects the statistical distributions of the classifier's similarity scores. The effect clearly hints on the existence of a causal link between the change of quality of signals and the alteration of the score distributions. If one succeeds to aptly capture the relevant signal qualities in the form of a quality measure, this causal relationship can be represented in the probabilistic terms as a statistical dependence between the relevant quality measure and the scores.

In this chapter we focus on the implications of this dependence. We discuss in detail why such dependence can be observed and we show how it can help to separate between the classes. As a consequence of the improved class separation, we present a new theoretical approach to classification with quality measures, $Q - stack$. Based on the concept of classifier stacking, we show how to construct a classifier ensemble that grants a higher classification performance than the baseline classifier that uses the similarity scores alone. We show how the presented classifier ensemble applies to classification in a single classifier, multiple classifier, and multimodal scenarios.

Finally we provide a generalization of the presented approach. We show that the attempts of using quality measures found in the prior art can be thought of as special cases of the generalized approach presented in this thesis.

3.2 Similarity scores and signal quality, revisited

Consider a problem of dichotomization, as introduced in Section 2.3, where observations represented by classification features $\mathbf{f}$ are assigned to class A or class B, based on classification scores x. Our goal is to best separate class A from class B using available classification scores and the available quality information qm, which will manifest itself in the strive to minimize the classification errors. For simplicity and without the loss of generality of the logic of the approach presented here, we now assume that the cost of an erroneous assignment of an observation to either class is equal.

In this section we focus on the decisions made based on the classification scores. We assume that the score distributions are given or known from an empirical estimation, but the nature of the signal, extracted features or the used baseline classifier are not discussed. This assumption will allow us to generalize the reasoning given here to classification using any type of dichotomizer that makes the decision based on similarity scores and their distributions. This generalization will therefore be valid for the vast majority of classifiers in use.

Consider two random processes (classes), A and B, which generate observations f, subjected to a baseline classifier. The yield of the classifier's discriminant function is similarity score x', distributed according to the probability density functions (pdf) $p(x'|A)$ and $p(x'|B)$. Let Figure 3.1 show an instance of such distributions. The classification decision is done by comparing the observation x' to a decision threshold τ'.

Now, consider that the observations are contaminated by noise. First, let us analyze why errors happen in the presence of noise. Consider a noise-generating process N which generates noise instances n according to a pdf $p(n)$. In general, $p(x'|A)$, $p(x'|B)$ and $p(n)$ are not explicitly known. The process N and the class-generating processes A and B interact in an arbitrary way, producing noisy observations x which are shifted in respect to the noiseless scores by Δx.

$$x = x' + \Delta x = \Phi(n, x') \tag{3.1}$$

The nature of the function $\Phi(n, x')$ does not need to be given explicitly, as it merely denotes that there exists some connection between the noise-related signal quality degradation and the observed noisy classification scores x. In the presence of noise, x' denotes the "clean" scores, i.e. scores that *would be* observed if there was no noise present, and x are the scores that are actually observed. Analogically, in the presence of noise $p(x'|A)$ and $p(x'|B)$ are hypothetical distributions of noise-free scores, while the actually observed scores are distributed according to $p(x|A)$ and $p(x|B)$. Since in the noisy situation it is $p(x|A)$ and $p(x|B)$ that are observed and used to build the baseline classifier, scores x are compared to a threshold τ in order to obtain the classification decisions. This situation is illustrated in Figure 3.1.

In this thesis we are concerned with the situation where the noise decreases the separation between the class-conditional score distributions, thus increasing the probability of committing a classification error - but one can imagine a function $x = \Phi(n, x')$ which would actually increase the class separation. Such a case would be of benefit for classification and would not require any further attention. In this thesis we are concerned with mitigating the detrimental impact of noise on the classification performance, however with the aim that the methods developed here must not actually harm classification performance should the noise be irrelevant. If the impact of noise is indeed decreasing the separation between the classes then it would seem like a good idea to recover analytically the underlying distributions of $p(x'|A)$ and $p(x'|B)$. This, however, is extremely hard if at all feasible considering that the nature of the function of Φ is unknown and may be arbitrarily complex. Recovering signals lost in noise has been shown to be feasible when particular simplifying assumptions about the noise are made [75]. We therefore need to seek a different method

of improving the classification performance in the presence of arbitrary noise.

Consider a signal generated by class A, whose hypothetical, noise-free score x'_0 would be correctly assigned to class A since $p(x'_0|A) > p(x'_0|B)$. Let us now assume that the same observation x'_0 is affected by noise n, resulting in a noisy observation x_0. Now, the classifier operates on the noisy signal, we therefore move to the classifier described by $p(x|A)$ and $p(x|B)$. The impact of noise n on x' manifests itself in a translation (shift) of the similarity score by Δx. If $\Delta x > (x'_0 - \tau)$ the observation x_0 will be falsely assigned class label B, since $p(x_0|A) < p(x_0|B)$, as it is shown in Figure 3.1. Since score x_0 depends on the value of the shift Δx, and since Δx depends of the noise n and the noise-free scores x'_0, it becomes evident that there exists a dependency between the noise n and the observed score x_0. In practical situations, one routinely does not have the comfort of knowing the actual distributions of $p(x'|A)$ and $p(x'|B)$ since the data collected is already contaminated by noise N. Instead, the class-conditional distributions of the noisy observations $p(x|A)$ and $p(x|B)$ can be estimated from the available data.

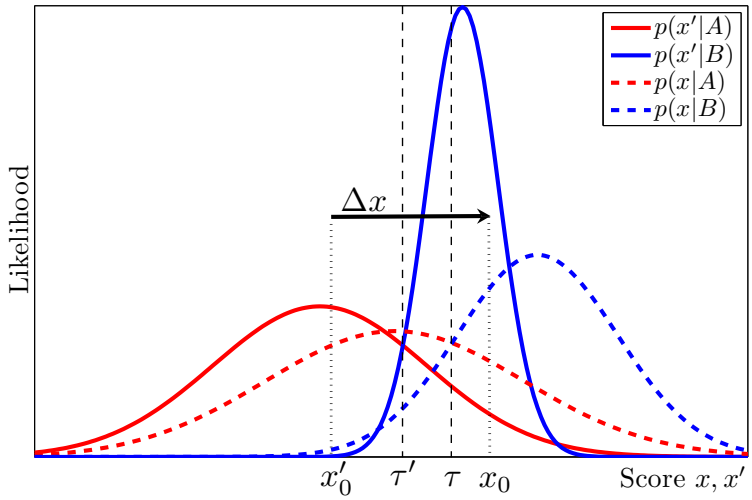


Figure 3.1: Distribution shift due to noise. In the presented example the distribution shift is caused by normally distributed additive noise.

As it is shown in Figure 3.1, the noise-affected class-conditional score distributions $p(x|A)$ and $p(x|B)$ can be altered in shape and translated in respect to the distributions of noise-free observations, $p(x'|A)$ and $p(x'|B)$. The type of distribution deformation and shift depends on the relationship between the noise N and the scores X'. Therefore if an alteration of the class-conditional distributions is observed, a statistical dependency between the noise and the similarity scores is evidently present. In the next sections we will show how this dependency can be exploited in order to improve classification.

Note that this model of interactions between noise and classification scores accounts as well for a situation where the observations are actually noise-free. In such situation $\Delta x = 0$, and $x_0 = x'_0$, which corresponds to the true state of nature. Therefore treating every observation as if they were potentially affected by noise is general and we adhere to it in this thesis.

The scores x can be observed directly since they are the immediate output of the baseline classifier before the decision is obtained by comparing the score to a preset threshold. However, the actual value of noise n that impacts x' is in general not obvious. Depending on the particular application

it may or it may not be feasible to actually measure the noise n itself. In certain cases it will be far more practical to assert the presence and impact of noise from the observed data itself. We shall refer to any measurement of the noise n, direct or indirect, as *quality measure*, which we will denote as qm. In the rest of this chapter and throughout the rest of the thesis the impact of noise will be considered through the prism of its measured impact on the observed data: the quality measures qm.

3.3 Classification with class-independent quality information

3.3.1 Evidence

As we have discussed in Section 3.2, in the presence of varying signal quality, the knowledge of the similarity score x is not sufficient to optimize the classification performance, or minimize the probability of committing a classification error. Considering Equation 3.1 a knowledge of the score x and the score shift is necessary in order to decrease the risk of misclassification to the risk that one is bound to run should the noiseless observations x be given. In other words, knowledge of x and n allows for modeling of the function Φ, and therefore also for an estimation of the shift of Δx. As discussed in Section 2.6, in many cases quality measures qm have to substitute the exact knowledge of n.

The similarity scores x and the quality measures are therefore complementary information relevant to the process of classification. Consequently, for every observation x we define the *evidence vector* e as a vector that consists of the similarity score and k quality measures relevant to this particular observation:

$$\mathbf{e} = [x, qm_1, qm_2, ..., qm_k]. \tag{3.2}$$

We stress here that the evidence vector, and therefore also the quality measures qm are pertinent to one particular observation x. For every new observation x a new evidence vector $\mathbf{e}$ is constructed.

3.3.2 Classification in the evidence space

In the classical paradigm of classification, the class label is assigned in the one-dimensional space of similarity scores x. Considering Equation 3.2 the classical paradigm is therefore equivalent to classification in the evidence space, where the evidence consists uniquely of scores $e = [x]$. Upon having quality measures $qm_{1,2,...,k}$ as additional components in the evidence vector the one-dimensional classification problem is expanded to become a $k+1$-dimensional problem, where k is the number of relevant quality measures involved.

In order to expect a benefit from additional dimensions in which the classification takes place it first needs to be ascertained that the class separation in the new evidence space is better than in the one-dimensional space of scores. If classes are to be separated in the space defined by evidence $\mathbf{e}$, obviously quality measures become classification features together with the scores x. It may not come through as an obvious thing how class-independent information - quality measures - may afford a better separation of classes. The key to understanding this seemingly counterintuitive effect is analyzing the impact of the fact that the components of the evidence vector are par excellence statistically dependent.

3.3.3 Dependence between features, feature relevance, and class separation

As we have discussed in detail in Section 2.6, quality measures qm constitute class-independent information. At the same time, as a part of the evidence vector $e = [x, qm]$ they become classification features, and so do the scores x. In feature selection literature, class-independent information is *individually irrelevant* to the given classification problem [38, 97]. However, individually irrelevant features can become relevant if used together with other features [67]. An example where two individually irrelevant features become relevant when used together is the well-known "checkerboard problem" [68]. The effect in which an individually irrelevant feature becomes relevant in the context of other features can be explained by a causal relationship between the features [66]. The dependence which exists between scores x and quality measures qm, which we have discussed in section 3.2, is an instance of such a causal relationship.

The problem of analyzing dependencies between features has been frequently simplified to analyzing the linear correlations between them. To some authors, non-zero correlation coefficient between features suggests redundancies in the class-selective information contained in the features [109]. Similar conclusions seem to be emerging from the work of Poh and Bengio [119]. At the same time, however, Fumera and Roli give some theoretical and practical justification to a quite opposite claim [53]. In [112] Nandakumar et al. analyze the impact of correlation between classification scores originating from multiple classifiers in the context of multimodal biometric fusion. Their conclusions suggest that while the correlation does have an impact on the fusion performance, this effect is negligible from a practical viewpoint. Further, Koval et al. look not only at correlation but, wider, at dependencies between features in multimodal fusion systems [84], and conclude that these dependencies actually help separate between classes. The authors provide an information-theoretic proof that dependencies between features indeed do reduce the bound on probabilities of misclassification, however, their conclusions seem to be too strong, and underlying assumptions not general enough, which we show later in this chapter. Nevertheless, the impact of dependencies between features on class separation did not pass unnoticed.

The issue of dependency and correlation between features and their impact on the class separation is indeed the central part of the approach presented in this thesis, it therefore deserves a closer attention in the following Sections. In particular, we aim to give the reader an intuition about why and under what conditions should mutually dependent features grant a better class separation. The reason to do so is twofold. First, an intuitive understanding of the impact of the dependencies between features on classification performance should give the reader a firm grip of the central concept of this thesis - how class-independent features can improve class-specific decision-making. Second, the reasoning we are about to unfold here is in direct and immediate link with the way we propose to apply presented findings in practical applications.

3.4 Towards an understanding of feature dependencies and their impact on classification

3.4.1 Theoretical foundations

Recall from Section 3.2 that we are interested in classification of observations in the evidence space, where the evidence is defined as a vector consisting of the actual classifier similarity scores x and quality measures qm. Here, for the clarity of the reasoning, we consider the use of one quality measure, thus for every observation the evidence vector $\mathbf{e} = [x, qm]$. It will become obvious later

how the generalization applies to different configurations of evidence.

The results of Koval et al. [84] analyze the impact of statistical dependencies on the probability of misclassification. However, it is arguably more difficult to demonstrate generally pairwise dependent data, then pairwise correlated data, which is a special case of dependence. Correlation necessarily entails dependency although dependent random variables may be uncorrelated [42, 82]. We will therefore here discuss the case of correlated data, where the correlation coefficient between dimensions for both classes is equal, and then extend the presented concepts to generally dependent data.

3.4.2 Bivariate normal case with equal correlation coefficients

Let us consider how correlation between features impacts class separation between classes A and B for bivariate normal class-dependent distributions of evidence $\mathbf{e} = [x, qm]$. For that we assume that class-conditional marginal distributions of scores $p(x|A)$, $p(x|B)$ and quality measures $p(qm|A)$, $p(qm|B)$ are normal. Without a loss of generality let us place the means of the $p(x|A)$ and $p(qm|A)$ at the origin of the axes. The marginal distributions are therefore given by:

$$\begin{aligned}
p\left(x|A\right) &= \mathcal{N}\left(\mu_{x,A}, \sigma_{x,A}\right) = \frac{1}{\sqrt{2\pi\sigma_{x,A}^2}} e^{-\frac{x^2}{2\sigma_{x,A}^2}} \\
p\left(x|B\right) &= \mathcal{N}\left(\mu_{x,B}, \sigma_{x,B}\right) = \frac{1}{\sqrt{2\pi\sigma_{x,B}^2}} e^{-\frac{\left(x-\mu_{x,B}\right)^2}{2\sigma_{x,B}^2}} \\
p\left(qm|A\right) &= \mathcal{N}\left(\mu_{qm,A}, \sigma_{qm,A}\right) = \frac{1}{\sqrt{2\pi\sigma_{qm,A}^2}} e^{-\frac{qm^2}{2\sigma_{qm,A}^2}} \\
p\left(qm|B\right) &= \mathcal{N}\left(\mu_{qm,B}, \sigma_{qm,B}\right) = \frac{1}{\sqrt{2\pi\sigma_{qm,B}^2}} e^{-\frac{\left(qm-\mu_{qm,B}\right)^2}{2\sigma_{qm,B}^2}},
\end{aligned} \tag{3.3}$$

where $\mu_{x,B}$ and $\mu_{qm,B}$ are the means of the marginal distributions of class B, and $\sigma_{x,A}^2$, $\sigma_{x,B}^2$, $\sigma_{qm,A}^2$ and $\sigma_{qm,B}^2$ are respective variances. Assume the Pearson's correlation coefficient between QM and X to be the same for both classes A and B. Let the value of the correlation coefficient be equal to ρ. Bivariate conditional distributions of $p(\mathbf{e} = [x, qm]|A)$ and $p(\mathbf{e} = [x, qm]|B)$ whose marginal distributions are normal are given by [82]:

$$\begin{aligned}
p(e[x, qm]|A) &= \frac{1}{2\pi\sigma_{x,A}\sigma_{qm,A}\sqrt{1-\rho^2}} e^{\left(-\frac{1}{2-2\rho^2}\left(\frac{x^2}{\sigma_{x,A}^2}+\frac{qm^2}{\sigma_{qm,A}^2}-\frac{2x\cdot qm}{\sigma_{x,A}\sigma_{qm,A}}\right)\right)} \\
p(e[x, qm]|B) &= \frac{1}{2\pi\sigma_{x,B}\sigma_{qm,B}\sqrt{1-\rho^2}} e^{\left(-\frac{1}{2-2\rho^2}\left(\frac{(x-\mu_{x,B})^2}{\sigma_{x,B}^2}+\frac{(qm-\mu_{qm,B})^2}{\sigma_{qm,B}^2}-\frac{2(x-\mu_{x,B})(qm-\mu_{qm,B})}{\sigma_{x,B}\sigma_{qm,B}}\right)\right)}.
\end{aligned} \tag{3.4}$$

In Equation 3.4 the correlation coefficient ρ is constrained to satisfy $-1 < \rho < 1$ in order to avoid a degenerate case where $x \propto qm$ [42]. Since we search for best possible class separation we again assume the costs of falsely assigning an observation to classes A and B equal. Considering this the

class separation can be expressed in terms of associated Bayes error [42] as given by Equation 3.5:

$$E_{Bayes} = 1 - \frac{1}{2}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} |p(\mathbf{e}|A) - p(\mathbf{e}|B)|\, dxdqm, \tag{3.5}$$

assuming equal prior class probabilities, as discussed in Section 2.3.1. Evaluating Equation 3.5 analytically is difficult and impractical. However, what we wish to show is not what precisely the associated Bayes error is, but rather to show how does the error depend on the correlation between qm and x. Since the error depends on the overlap between the class-conditional distribution, then the measure of error is given by

$$E = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \sqrt{p(\mathbf{e}|A)p(\mathbf{e}|B)}dxdqm, \tag{3.6}$$

which is known as the Matusita distance between distributions $p(\mathbf{e}|A)$ and $p(\mathbf{e}|B)$ [24, 169]. The intuition behind the formulation of Equation 3.6 is quite straightforward: if both $p(\mathbf{e}|A)$ and $p(\mathbf{e}|B)$ are non-overlapping then $p(\mathbf{e}|A)p(\mathbf{e}|B) = 0$ for every instance of e, and therefore the error measure is minimized at $E = 0$. In the opposite case the maximal overlap happens for $p(\mathbf{e}|A) = p(\mathbf{e}|B)$ for every instance of e, and in this case the error measure is maximized and yields $E = 1$ [110]. The relevance of thus defined error measure is demonstrated in figures 3.4 and 3.5.

Now we proceed to show how the measure of error E depends on the correlation between x and qm for the case when $p(\mathbf{e}|A)$ and $p(\mathbf{e}|B)$ are bivariate normal distributions and correlation coefficient ρ is equal for both classes. For simplicity here we assume that $\sigma^2_{qm,A} = \sigma^2_{qm,B} = \sigma^2_{qm}$ and $\sigma^2_{x,A} = \sigma^2_{x,B} = \sigma^2_x$. In this case Equation 3.6 takes following shape.

$$E = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \sqrt{\frac{e^{\left(-\frac{x^2}{\sigma_x^2} - \frac{2\rho x qm}{\sigma_x\sigma_{qm}} + \frac{qm^2}{\sigma_{qm}^2}\right)} e^{\left(-\frac{(x-\mu_x)^2}{\sigma_x^2} - \frac{2\rho(x-\mu_x)(qm-\mu_{qm})}{\sigma_x\sigma_{qm}} + \frac{(qm-\mu_{qm})^2}{\sigma_{qm}^2}\right)}}{4\pi^2\sigma_x^2\sigma_{qm}^2(1-\rho^2)}} dxdqm \tag{3.7}$$

Equation 3.7 evaluates to:

$$E = e^{\frac{x^2\sigma_{qm}^2 + qm^2\sigma_x^2 - 2\rho\mu_x\mu_{qm}\sigma_x\sigma_{qm}}{8\sigma_x^2\sigma_{qm}^2(\rho^2-1)}} \tag{3.8}$$

In order to find and identify the values of ρ for which $E(\rho)$ may reach an extremum we find ρ_0 that satisfies $\frac{\sigma E(\rho)}{\sigma\rho} = 0$:

$$\begin{aligned}
&\frac{dE}{d\rho} = \frac{d}{d\rho} e^{\frac{\mu_x^2\sigma_{qm}^2 + \mu_{qm}^2\sigma_x^2 - 2\rho\mu_x\mu_{qm}\sigma_x\sigma_{qm}}{8\sigma_x^2\sigma_{qm}^2(\rho^2-1)}} = 0 \Rightarrow \\
&\Rightarrow -\frac{1}{4} e^{\frac{\mu_x^2\sigma_{qm}^2 + \mu_{qm}^2\sigma_x^2 - 2\rho\mu_x\mu_{qm}\sigma_x\sigma_{qm}}{8\sigma_x^2\sigma_{qm}^2(\rho^2-1)}} \left(\frac{\rho\left(\mu_x^2\sigma_{qm}^2 + \mu_{qm}^2\sigma_x^2\right) - \mu_x\mu_{qm}\sigma_x\sigma_{qm}\left(\rho^2+1\right)}{\sigma_x^2\sigma_{qm}^2\left(\rho^2-1\right)^2}\right) = 0 \Rightarrow \\
&\Rightarrow \rho\left(\mu_x^2\sigma_{qm}^2 + \mu_{qm}^2\sigma_x^2\right) - \mu_x\mu_{qm}\sigma_x\sigma_{qm}\left(\rho^2+1\right) = 0 \Rightarrow \\
&\Rightarrow \left(\rho\mu_{qm}\sigma_x - \mu_x\sigma_{qm}\right)\left(\rho\mu_x\sigma_{qm} - \mu_{qm}\sigma_x\right) = 0
\end{aligned} \tag{3.9}$$

Equation 3.9 holds for:

$$\rho_1 = \frac{\mu_{qm}\sigma_x}{\mu_x\sigma_{qm}}, \rho_2 = \frac{\mu_x\sigma_{qm}}{\mu_{qm}\sigma_x} \tag{3.10}$$

It is easy to notice that for $\mu_x \neq 0$ and $\mu_q m \neq 0$, $\rho_1 = \frac{1}{\rho_2}$. The correlation coefficient is by default constrained to $-1 > \rho > 1$ [42] therefore only one solution is a valid one for a given set of parameters of processes A and B, and therefore $E(\rho)$ has only one stationary point within the valid range of ρ.

We examine the nature of this stationary point for $\rho_1 = \frac{\mu_{qm}\sigma_x}{\mu_x\sigma_{qm}}$. The second derivative of $E(\rho)$ evaluates at point ρ_1 to:

$$\frac{d^2E}{d\rho^2} = \frac{1}{4}e^{\frac{\mu_x^2\sigma_{qm}^2 - \mu_{qm}^2\sigma_x^2}{8\sigma_x^2\sigma_{qm}^2\left(\left(\frac{\mu_{qm}\sigma_x}{\mu_x\sigma_{qm}}\right)^2 - 1\right)}} \frac{\mu_x^2\sigma_{qm}^4}{\sigma_x^2(\mu_{qm}^2\sigma_x^2 - \mu_x^2\sigma_{qm}^2)} \tag{3.11}$$

Note that in Equation 3.11

$$\frac{1}{4}e^{\frac{\mu_x^2\sigma_{qm}^2 - \mu_{qm}^2\sigma_x^2}{8\sigma_x^2\sigma_{qm}^2\left(\left(\frac{\mu_{qm}\sigma_x}{\mu_x\sigma_{qm}}\right)^2 - 1\right)}} \frac{\mu_x^2\sigma_{qm}^4}{\sigma_x^2} > 0$$

for all $\mu_x \neq 0$. This condition can be reasonably assumed to be always met: if $\mu_x = 0$ then classes $p(\mathbf{e}|A) = p(\mathbf{e}|B)$ and no classification is possible whatsoever. Since $-1 < \rho_1 < 1$ then necessarily

$$\mu_{qm}^2\sigma_x^2 < \mu_x^2\sigma_{qm}^2 \Rightarrow \frac{d^2E}{d\rho^2} < 0 \tag{3.12}$$

and consequently $E(\rho)$ has a maximum at ρ_1. If ρ_2 is assumed to satisfy $-1 < \rho_2 < 1$ then identically it follows that $E(\rho)$ has a maximum at ρ_2 since the difference in 3.12 will be merely symbolic - reversed indices for μ and σ.

We have shown in Equations 3.9 and 3.12 that the bound of Bayes error for Gaussian processes A and B, whose marginal distributions are fixed, depends on the correlation coefficient ρ between scores x and quality measures qm, and if $p(\mathbf{e}|A) \neq p(\mathbf{e}|B)$ there exists one and only one value of ρ where this error is maximal.

Let us now return to the discussed problem of classification in the evidence space. As defined, the dimension of qm does not provide in itself any class-separation, in other words the respective class-conditional marginal distributions are identical $p(qm|A) = p(qm|B)$. Since we had placed $p(qm|A)$ at the origin of the coordinate system, and $\mu_{qm} = 0$. The value of ρ that maximizes $E(\rho)$ is thus the valid solution to Equation 3.9 yields

$$\rho_1 = \frac{\mu_{qm}\sigma_x}{\mu_x\sigma_{qm}} = 0. \tag{3.13}$$

In Appendix A the same result is derived from the information-theoretical viewpoint, by analyzing the relative divergence between distributions A and B.

Based on Equation 3.13 and the above proof we can formulate following statement: **The Bayes error for two bivariate Gaussian processes, whose class-conditional marginal distributions are identical in one of the dimensions, is maximal if the dimensions are not correlated, given equal correlation coefficients between dimensions for both classes.**

Here we have given an analytical proof for a specific case when $\sigma_{qm,A}^2 = \sigma_{qm,B}^2 = \sigma_q m^2$. Of course this case is a special one. It can be also analytically proven that given the same assumptions about equality of correlation coefficients in both classes, for $\sigma_{qm,A}^2 \neq \sigma_{qm,B}^2$ the divergence between two Gaussian distributions is minimal for one and only one correlation coefficient $-1 < \rho < 1$. However, this proof is of limited relevance to the main topic of this thesis, and considering its significant notational size and complexity we prefer to skip it. Instead, we chose to show a series of numerically evaluated examples together with their graphical representations in Appendix A, which in our opinion is more illustrative and gives the reader a better intuitive understanding of the findings of this thesis and their implications.

3.4.3 Graphical representations and numerical simulations

In this section we represent and interpret the results from section 3.4.1 graphically in order to give the reader a better intuitive grasp of the key concepts of this thesis. Like in Section 3.4.1, consider two processes A and B described by bivariate Gaussian distributions $p(\mathbf{e} = [x, qm]|A)$ and $p(\mathbf{e} = [x, qm]|B)$. Recall that x is the dimension defined by the similarity scores returned by a classifier and qm is the dimension defined by relevant quality measure. By definition, the class-conditional marginal distributions of qm are equal, $p(qm|A) = p(qm|B)$. This situation is depicted

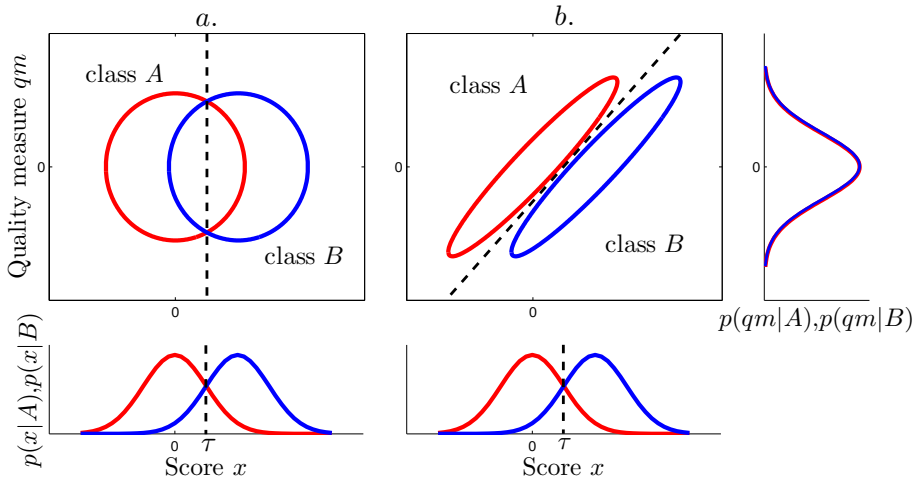
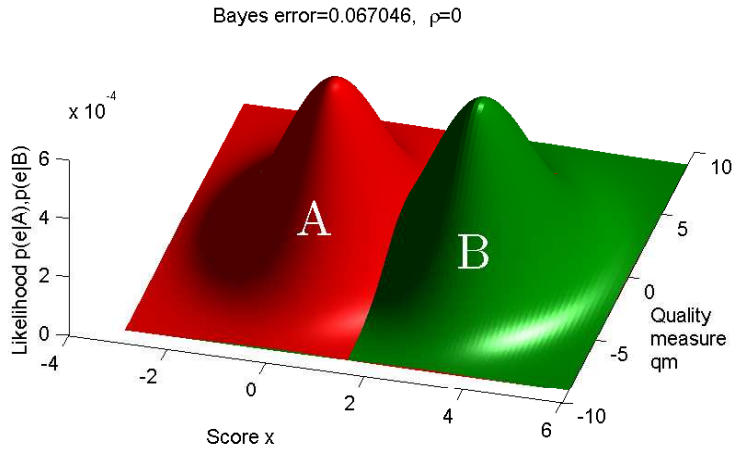


Figure 3.2: Improved class separation in the evidence space $e = [x, qm]$ due to the non-zero correlation ρ between scores x and quality measures qm. Plot (a) shows the case when $\rho = 0$, plot (b) shows the case for $\rho > 0$. The class-conditional marginal distributions of x and qm are shown. Ellipses drawn for $p(\mathbf{e}|A) = p(\mathbf{e}|B) = \varsigma$ where $\varsigma \approx 3\sigma^2_{x,A}$. The value of ς is identical in plots (a) and (b)

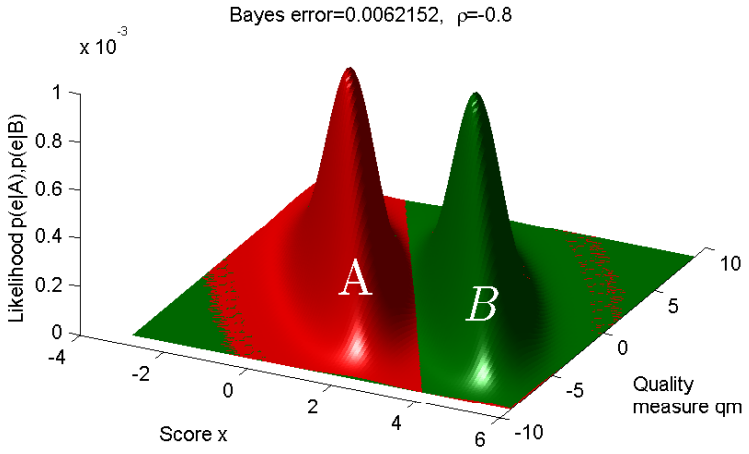
In Figure 3.2(a) the correlation between scores x and quality measures qm is $\rho = 0$, the data is thus uncorrelated. Since presented example deals with a Gaussian process, $\rho = 0$ entails independence of x and qm. In Figure 3.2(b) the correlation coefficient $\rho \neq 0$. Note that the class-conditional marginal distributions are identical in both subplots (a) and (b) in Figure 3.2. The ellipses are level curves drawn for $p(\mathbf{e}|A) = p(\mathbf{e}|B) = \varsigma$ where $\varsigma \approx 3\sigma^2_{x,A}$. The value of ς is immaterial and serves the demonstration purpose only as long as is kept constant across figure 3.2(a) and (b). The distance between the ellipses and the amount of area enclosed by the intersection of the two ellipses in each subplot correspond to the actual separation between classes A and B. The further apart the ellipses the better class separation and consequently the lower the associated Bayes error.

Note that in spite of the fact that the marginal class-conditional distributions are identical in both Figure 3.2(a) and (b), the class separation is much better in the case (b) than in case (a). This difference is due only to the existing non-zero correlation between scores x and quality measures qm, which is a manifestation of their causal link discussed in Secition 3.2. The dashed lines correspond to the optimal decision boundaries separating classes A and B in the evidence space. In Figure 3.2(a) this boundary is a line parallel to the qm axis and therefore it provides identical class separation as the marginal distributions $p(x|A)$ and $p(x|B)$ alone. In this situation the quality measures qm are irrelevant to the classification task and consequently are not useful at all in the classification process. In the subplot (b) the decision boundary could not be replicated in the score domain alone.

Therefore we can conclude that if the correlation between classes score x and qm is $\rho \neq 0$ the quality measures provide an additional degree of freedom in which classes can be separated better. As a consequence of the causal relationship between scores x and quality measures qm, the quality information becomes a relevant classification feature. This graphical representation fully reflects the analytical results from section 3.4.1 and Appendix A.



(a) $\rho = 0$



(b) $\rho = -0.8$

Figure 3.3: A three-dimensional representation of the gain in separation between classes A and B due to the interdimensional feature correlation. Correlation coefficient (a): $\rho = 0$, (b): $\rho = -0.8$.

Figure 3.3 shows this very same effect using a pseudo-three-dimensional representation of distributions $p(\mathbf{e}|A)$ and $p(\mathbf{e}|B)$. In this particular example the parameters of $p(\mathbf{e}|A)$ and $p(\mathbf{e}|B)$ are following: $\mu_{x,A} = \mu qm, A = \mu qm, B = 0$, $\sigma^2_{x,A} = \sigma^2_{x,B} = 1$, $\sigma^2_{qm,A} = \sigma^2_{qm,B} = 3$, $\mu_{qm,B} = 3$. In Figure 3.3(a) the correlation coefficient $\rho = 0$ and in (b) $\rho = -0.8$. The improved class separation in Figure 3.3(b) over the case presented in Figure 3.3(a) is readily visible. The figure also quotes the corresponding numerical approximations of the Bayes error. The class separation gain due to the change in correlation between x and qm is approximately one order of magnitude.

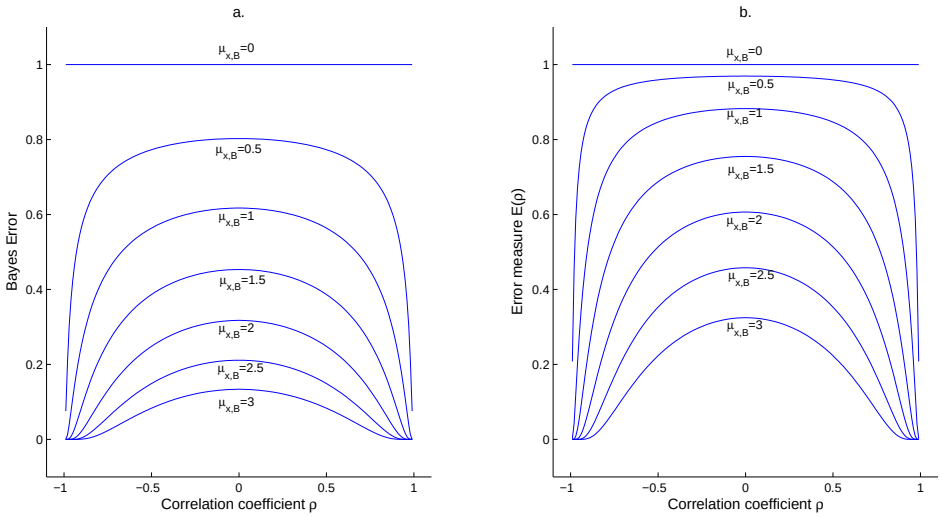


Figure 3.4: Bayes error (a) and error measure $E(\rho)$ (b) defined by Equation 3.6 as a function of correlation coefficient ρ. Curves drawn for varying values of $\mu_{x,B}$

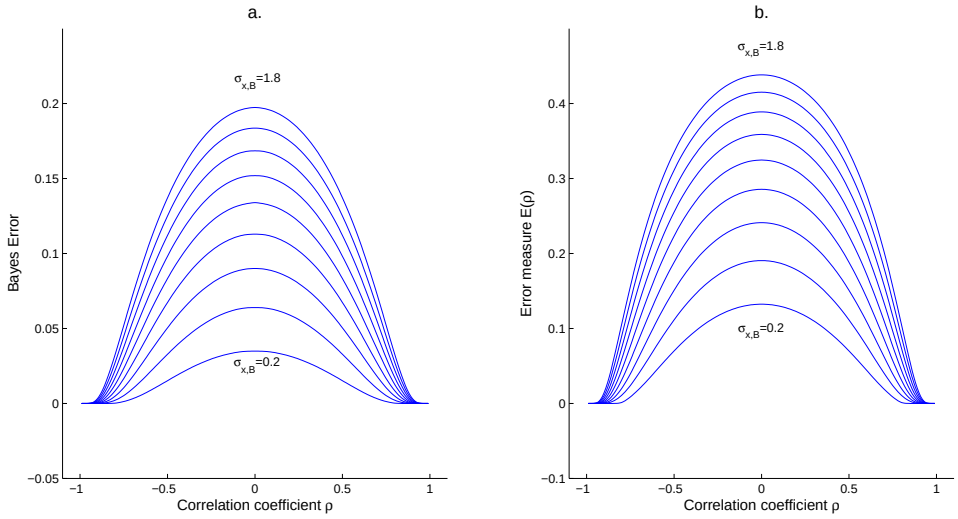


Figure 3.5: Bayes error (a) and error measure $E(\rho)$ (b) defined by Equation 3.6 as a function of correlation coefficient ρ. Curves drawn for varying values of $\sigma_{x,B}$

The same base parameters for processes A and B have been used in computing the curves in Figure 3.4 in order to demonstrate the actual influence of the correlation coefficient ρ on the class separation. Here we quantify the class separation by computing the numerical approximation of the associated Bayes error for the full possible range of values $-1 < \rho < 1$. Next to the approximated Bayes error in Figure 3.4 (a) we provide corresponding curves for the Matusita distance computed by Equation 3.6 - Figure 3.4 (b). In both cases we show the error curves as the function of ρ computed for a row of values of $\mu_{x,B}$, which corresponds to 'pushing away' distribution B from A in the dimension of scores x. Note that the shape of the subsequent curves is similar for subfigures (a) and (b) and in both cases the shape changes in a similar way with the change of $\mu_{x,B}$, which shows the appropriateness of the chosen distance measure discussed in Section 3.4.1. Indeed, as expected from the theoretical proofs from Section 3.4.1 and Appendix A, for fixed parameters of marginal distributions the class separation in the evidence space is minimal (and associated Bayes error maximal) for independent (uncorrelated, $\rho = 0$) scores x and quality measures qm.

The theoretical derivations from Section 3.4.1 and Appendix A do not include the case where $\sigma_{x,A} \neq \sigma_{x,B}$. The numerically evaluated error curves that demonstrate this case are shown in Figure 3.5. Here, $\mu_{x,B} = 3$ and $0.2 \leq \sigma_{x,B} \leq 1.8$ while other parameters are still $\mu_{x,A} = \mu qm, A = \mu qm, B = 0$, $\sigma_{x,A} = 1$, $\sigma_{qm,A} = \sigma_{qm,B} = 3$. Again, changing the value of $\sigma_{x,B}$ did not change the recurring effect that correlation between scores x and quality measures qm allows for better class separation than it is the case for uncorrelated dimensions.

Since the main topic of this thesis is the use of class-independent quality measures in order to improve classification, we do not discuss here the cases where $\mu_{qm,B} \neq 0$ and $\sigma_{qm,A} \neq \sigma_{qm,B}$. It must be however noted that these cases are highly interesting from the viewpoint of general theory of pattern classification. We therefore treat them in more detail and represent graphically in Appendix A.

3.4.4 Extension to more than two dimensions

The reasoning presented in Section 3.4.1 was explicitly considering a two-dimensional case, in which one of the dimensions (scores x) was class-selective and the other one (quality measures qm) did not in itself provide any class-separation. In this section we show how the results from Section 3.4.1 extend their validity to more than two dimensions.

Assume that classes A and B have joint distributions $p(x, qm, y|A)$ and $p(x, qm, y|B)$ in three evidence dimensions. Their level volumes $p(x, qm, y|A) = p(x, qm, y|B) = const.$ are shown in Figure 3.6. In top of the two dimensions of x and qm that the reader is already familiar with, we have added a third dimension y. For the moment we do not specify if dimension y provides by itself class-specific information, and if y depends or not on x or qm. The evidence space is now defined by triplets of evidence $\mathbf{e} = [x, qm, y]$.

Let us consider the class separation in the new, three-dimensional evidence space, and the corresponding Bayes error E''_{Bayes} in the new evidence space, again assuming equal error cost for erroneous assignment of an observation to A or B. The Bayes error E''_{Bayes} can be conveniently expressed in terms of the Kolmogorov variational distance [40] by:

$$E''_{Bayes} = \frac{1}{2} - \frac{1}{2} \int_y \int_{qm} \int_x |p(x, qm, y|A) - p(x, qm, y|B)| dx dqm dy = \int_y E_{Bayes}(y) dy, \qquad (3.14)$$

where E_{Bayes} is the Bayes error between the two-dimensional distributions $p(x, qm, y = y_0|A)$ and $p(x, qm, y = y_0|B)$. Note that $p(x, qm, y = y_0|A)$ and $p(x, qm, y = y_0|B)$ are equal to marginal distributions $p(x, qm|A)$ and $p(x, qm|B)$ if and only if y is not correlated with neither x nor y. A

graphical representation of $p(x, qm, y = y_0|A)$ and $p(x, qm, y = y_0|B)$ is shown in Figure 3.6. Note that the distributions $p(x, qm, y = y_0|A)$ and $p(x, qm, y = y_0|B)$ are defined on a plane orthogonal to the axis y, intersecting $p(x, qm, y|A)$ and $p(x, qm, y|B)$, and defined by $y = y_0$. The class boundary between $p(x, qm, y = y_0|A)$ and $p(x, qm, y = y_0|B)$ is depicted as a solid line.

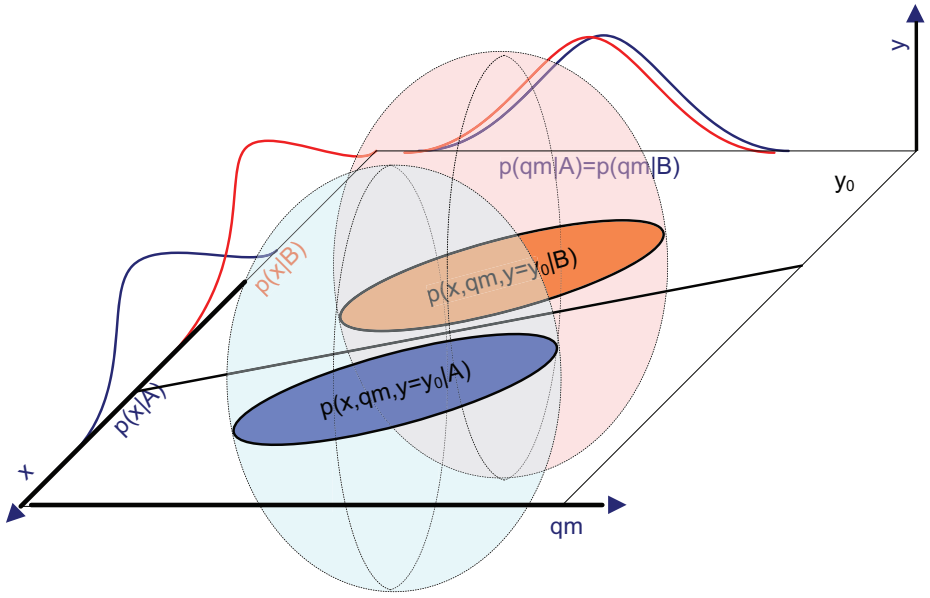


Figure 3.6: Improved separation in the 3-dimensional evidence space $\mathbf{e} = [x, qm, y]$

For every value of y_0 the intersection of the plane $y = y_0$ with the three-dimensional distributions $p(x, qm, y|A)$ and $p(x, qm, y|B)$ will produce different $p(x, qm, y = y_0|A)$ and $p(x, qm, y = y_0|B)$ and therefore also corresponding error E_{Bayes} depends on y - notation $E_{Bayes}(y)$ is therefore justified.

For each $y = y_0$, distributions $p(x, qm, y = y_0|A)$ and $p(x, qm, y = y_0|B)$ behave as discussed in Section 3.4.1 and $E_{Bayes}(y_0)$ depends on the correlation ρ between x and qm. Consequently, if $\rho = 0$ then $E_{Bayes}(y_0)$ will be maximal.

Since we made no assumptions on y whatsoever, we cannot also assume that ρ will remain constant for all y. However, as a consequence of Equation 3.14 E''_{Bayes} will be maximal if and only if $\bigwedge_{y_0 \in (-\infty,\infty)} \rho(y_0) = 0$. In other words, non-zero correlation between x and qm that increases the class separation between $p(x, qm, y = y_0|A)$ and $p(x, qm, y = y_0|B)$ will necessarily also increase class separation in the three-dimensional space $\mathbf{e} = [x, qm, y]$.

In the above reasoning we made no assumptions as to the relationships between y and the remaining dimensions of x and qm. Adding y could therefore only improve the theoretical class separation. Note that should y be also dependent on x, by the same reasoning as above class separation would be further improved. Adding an extra dimension that is independent from both x and qm is of no benefit or interest. Therefore we reach an important conclusion that **in order to obtain improved separation between classes, the components of the feature vector must be pair-wise dependent**. Obviously, at least one of the dimensions must provide saparation between classes. Remaining components of the evidence vector may be either class-selective (and therefore by definition dependent on scores x), or class-independent but otherwise dependent on scores x.

3.4.5 Non-Gaussian distributions

In some practical applications the assumption of Gaussianity of distribution is not far from reality and it simplifies the analytical analysis [119]. However it may not always be a good assumption and in such situations it is important to extend the findings presented in Section 3.4.1 to non-Gaussian cases.

It is a well-known fact that any non-Gaussian distribution can be approximated up to an arbitrary accuracy by a linear mixture of Gaussian distributions - Gaussian Mixture Models [111, 126]:

$$p(\mathbf{e}|A) = \sum_i \mathcal{N}(w_i, \mu_i, \sigma_i^2), \;\; p(\mathbf{e}|B) = \sum_j \mathcal{N}(w_j, \mu_j, \sigma_j^2), \tag{3.15}$$

where in general case $i \neq j$.

This situation is depicted in Figure 3.7, where two paired components of arbitrary distributions $p(\mathbf{e}|A) = p_1(\mathbf{e}|A) + p_2(\mathbf{e}|A)$ and $p(\mathbf{e}|B) = p_1(\mathbf{e}|B) + p_2(\mathbf{e}|B)$. The correlation coefficients for the component pairs are respectively ρ_1 and ρ_2, $\rho_1 \neq \rho_2$. Note that classes A and B can be better separated for $\rho_1 \neq 0, \rho_2 \neq 0$ than it would be the case if $\rho_1 = \rho_2 = 0$, which is represented by the decision boundary $x = \tau$.

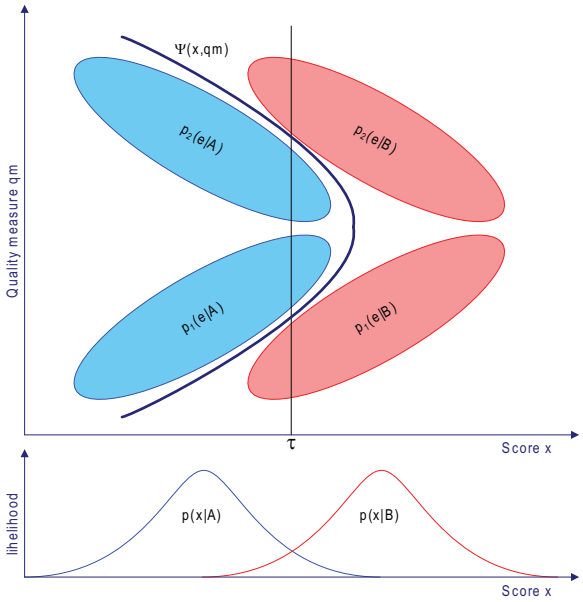


Figure 3.7: Improved separation in the evidence space $e = \{x, qm\}$ for a Gaussian mixture model (GMM).

The reader can easily recognize from Figure 3.7 that if both classes A and B are represented by a two-component Gaussian Mixture Model then their associated minimal Bayes error will be a function of the parameters w, μ and σ for each component, and of the correlation coefficients between x and qm. There is no reason to universally assume the correlation coefficient ρ to be equal for every component of the mixture, as it is depicted in Figure 3.7. In the next section we discuss the impact of different correlation coefficients in each class on class separation in the Gaussian case.

3.4.6 Bivariate normal case with unequal correlation coefficients

So far we have been making an explicit assumption that the correlation coefficient between the evidence components for both classes is the same, which is also a tacit assumption of the results

given in [84]. In this section we consider the case when this assumption is violated. Note that the class-independence of the quality measures does not entail the class independence of the correlation coefficient between evidence components because the dependence relation is defined by both the class-independent noise and the class-dependent classification scores. Although the correlation between quality measures and scores may be class-dependent, the main premise of this thesis that quality measures are not class-selective by themselves is not invalidated.

Let us now return to the bivariate normal example discussed earlier in this chapter. Again we shall use the very same marginal distributions of evidence as given by Equation 3.3. Now, the normal bi-variate joint class-conditional evidence distributions are given by

$$\begin{aligned} p(e(x,qm)|A) &= \frac{1}{2\pi\sigma_{x,A}\sigma_{qm}\sqrt{1-\rho_A^2}} e^{\left(-\frac{1}{2-2\rho_A^2}\left(\frac{x^2}{\sigma_{x,A}^2}+\frac{qm^2}{\sigma_{qm}^2}-\frac{2x\cdot qm}{\sigma_{x,A}\sigma_{qm}}\right)\right)} \\ p(e(x,qm)|B) &= \frac{1}{2\pi\sigma_{x,B}\sigma_{qm}\sqrt{1-\rho_B^2}} e^{\left(-\frac{1}{2-2\rho_B^2}\left(\frac{(x-\mu_x)^2}{\sigma_{x,B}^2}+\frac{qm^2}{\sigma_{qm}^2}-\frac{2(x-\mu_x)qm}{\sigma_{x,B}\sigma_{qm}}\right)\right)}, \end{aligned} \tag{3.16}$$

where ρ_A and ρ_A are the correlation coefficients between scores x and quality measures qm for class A and B, respectively. The Matusita distance E between $p(e(x,qm)|A)$ and $p(e(x,qm)|B)$ computed again according to Equation 3.6 yields

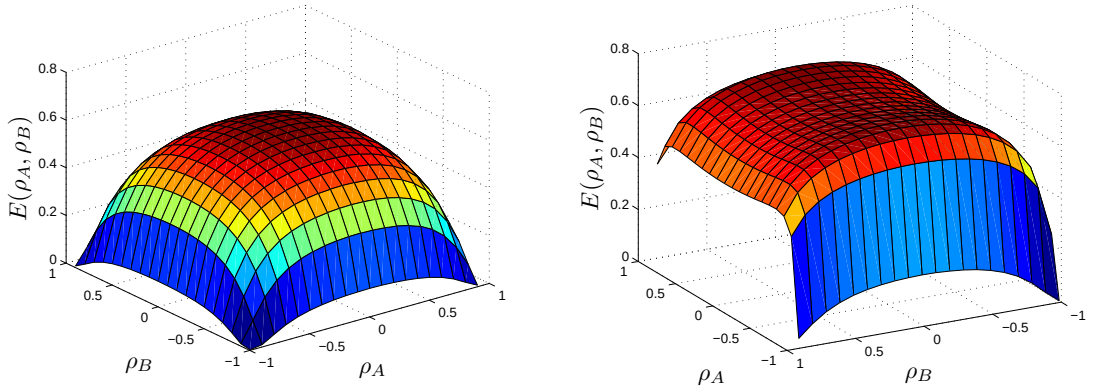
$$E = -\frac{2e^{\frac{\mu_x}{2\zeta}}\sqrt{\sigma_{x,A}\sigma_{x,B}}(\rho_A^2-1)^{\frac{1}{4}}(\rho_B^2-1)^{\frac{1}{4}}}{\zeta}, \tag{3.17}$$

where $\zeta = (\sigma_{x,A}\rho_A + \sigma_{x,B}\rho_B)^2 - 2(\sigma_{x,A}^2 + \sigma_{x,B}^2)$. Solving the set of equations $\{\frac{\delta}{\delta\rho_A}E(\rho_A,\rho_B) = 0, \frac{\delta}{\delta\rho_B}E(\rho_A,\rho_B) = 0\}$ for ρ_A and ρ_B yields two solutions

$$\{\rho_A = \rho_B = 0\}, \{\frac{\sigma_{x,A}^2}{\sigma_{x,B}^2} = \frac{\rho_A^2-1}{\rho_B^2-1}\}. \tag{3.18}$$

Note that the results of optimization of $E(\rho_A,\rho_B)$ $\rho_A = \rho_B$ presented in Section 3.4.2 are accommodated by the more general results given by Equation 3.18. However, for $\rho_A \neq \rho_B$ function $E(\rho_A,\rho_B)$ can have either an extremum or an infliction point, depending on the actual values of ρ_A, ρ_B, $\sigma_{x,A}$ and $\sigma_{x,B}$. The detailed analysis of the conditions which must be met for $E(\rho_A,\rho_B)$ to reach an extremum at any of the loci defined by Equation 3.18 is rather tedious and is not of direct relevance here and therefore we decide to skip it. What is of importance, however, is to realize that for $\rho_A \neq \rho_B$ the link between the class separation and the correlations between evidence components is a complex one. This fact is illustrated by an example in Figure 3.8, where the impact of changing variance $\sigma_{x,B} = 10$ changes the location and type of stationary points of the Matusita distance between two bivariate normal distributions.

It is a good moment to stop the theoretical derivations and consider the practical implications of these findings. We have derived a relationship between parameters of bivariate normal distributions which define the stationary points of a distance function that describes the overlap between the joint class-conditional distributions of evidence. This result could be directly applicable under the assumption that the data we encounter in practice is normally distributed in the evidence space. It is often not the case. In particular, normal data distribution assumes that all evidence components can take any values from the range of $(-\infty,\infty)$. This assumption, which extends to GMM modeling, is often contradicted by the physical or mathematical constraints specific to the data under consideration. For multivariate non-Gaussian processes the computation of the integrals necessary in deriving the analytical distance measures becomes very complicated and impractical, and analytical derivations are replaced by numerical approximations.



(a) $\sigma_{x,A} = \sigma_{x,B} = 1, \mu_x = 2$

(b) $\sigma_{x,A} = 1, \sigma_{x,B} = 10, \mu_x = 2$

Figure 3.8: Impact of the relationships between variances $\sigma_{x,A}$, $\sigma_{x,B}$ and correlation coefficients ρ_A, ρ_B on the location of stationary points the Matusita distance $E(\rho_A, \rho_B)$ between two bivariate normal distributions.

The same implications are valid for the problem of classification with class-independent quality information, which is the subject of this thesis. Instead of trying to predict the class separation from the correlation estimates, which is likely to introduce errors due to violations of the Gaussianity constraint, we postulate a data-driven approach where the impact of dependencies between evidence components on class separation is learned from available training data.

3.4.7 Data-driven statistical classification in the evidence space

So far we have been considering the class-conditional distributions which were either Gaussian, or could be represented with a GMM. As we have discussed, the Gaussianity assumption is often inappropriate. The GMM models used in certain practical applications in biometrics can be made up of a very large number of components (for example 512 GMM components for face verification in [137]). In this situation analytical derivations of the theoretical Bayes error would be utterly impractical. Moreover, in reality representing $p(\mathbf{e}|A)$ and $p(\mathbf{e}|B)$ by a GMM may not be the best choice of representation and a non-parametric, histogram-based methods of estimating these distributions may be used. In this case analytical derivations of the impact of dependence between evidence components on class separation are simply not possible.

As a consequence of considering normally distributed data, so far we have been using the notion of correlation between the dimensions in lieu of dependency, which is not equivalent, albeit far easier to handle in algebraic notation and representation. In general, the dimensions are not always globally correlated if dependent, and the type of dependency is not necessarily the same for both classes A and B. For these reasons the examples presented earlier in this chapter are merely special cases of class separation in the evidence space of mutually dependent dimensions. Yet they are illustrative examples. Following their logic, here we will show how inter-dimensional feature dependencies relate to correlations and to class separation for distributions of arbitrary shapes.

In Section 3.4.1 we have used an exact notion of the underlying statistical processes that generate data from classes A and B in order to show how classifying data in the evidence space is affected by the correlation between the components of the evidence vector $e = [qm, x]$. In practice, however,

one rarely has the comfort of knowing the analytical representations of these processes. The same is true for the quality measures: the function that links available quality measures with the scores may be difficult to infer analytically. However, recall from Section 2.2 that in a typical statistical pattern classification problem one has got a *training* data set S_{TR} at hand. If data set S_{TR} is representative for the population from which the testing set S_{TS} is drawn then a classifier that best classifies S_{TR} will retain its properties when applied to classify S_{TS}.

Consider the separation between distributions shown in Figures 3.2 and 3.7. Note that in both cases the separation between classes A and B was improved in the evidence space in respect to the classification using marginals $p(x, A)$ and $p(x|B)$, which manifested itself in the deviation of the decision boundary $\Psi(x, qm)$ from the decision boundary $x = \tau$, where τ is the decision threshold in the score domain. As a consequence - by observing the behavior of the decision boundary $\Psi(x, qm)$ in the evidence space and comparing it to $x = \tau$ for the training set S_{TR} it is possible to assess the benefits of using quality measures qm as a component of the evidence vector.

Let us demonstrate this on two practical examples. In both examples, 10000 data points are generated by a linear mixture of Gaussian processes. Each of these processes is defined in the two-dimensional evidence space $\mathbf{e} = [x, qm]$. The class-independence of marginal distributions $p(qm|A)$ and $p(qm|B)$ is maintained. The knowledge of these processes or their parameters is not used in the classification process. The data is then classified in three ways:

1. Thresholding in the score domain, $e = [x]$, with the decision boundary $x = \tau$. The threshold value τ is minimizing the total error for classification of $x|A$ and $x|B$.

2. Bayes classifier using GMM operating in the evidence space $\mathbf{e} = [x, qm]$. Decision boundary $\Psi_{Bayes}(x, qm)$ is defined in the evidence space. Parameters of the GMM class models estimated using the EM algorithm [42].

3. SVM classifier operating in the evidence space $\mathbf{e} = [x, qm]$. Decision boundary $\Psi_{SVM}(x, qm)$ is defined in the evidence space.

Example 1: Linearly separable case

In this example we show how improved class separation and consequently classification can be achieved in the evidence space of $e = [x, qm]$ for linearly separable classes. Consider data generated by a linear mixture of two Gaussian processes defined by parameters:

$$\Sigma_{A,1} = \begin{pmatrix} 0.5 & 0.095 \\ 0.095 & 0.2 \end{pmatrix}, \Sigma_{A,2} = \begin{pmatrix} 0.4 & 0.04 \\ 0.04 & 0.1 \end{pmatrix}, \mu_{\mathbf{A,1}} = \begin{pmatrix} 3 & 2 \end{pmatrix}^T, \mu_{\mathbf{A,2}} = \begin{pmatrix} 0 & 0 \end{pmatrix}^T.$$

$$\Sigma_{B,1} = \begin{pmatrix} 0.5 & 0.063 \\ 0.063 & 0.8 \end{pmatrix}, \Sigma_{B,2} = \begin{pmatrix} 0.4 & -0.18 \\ -0.18 & 0.9 \end{pmatrix}, \mu_{\mathbf{B,1}} = \begin{pmatrix} 3 & -1 \end{pmatrix}^T, \mu_{\mathbf{B,2}} = \begin{pmatrix} 0 & 1 \end{pmatrix}^T.$$

The data distribution (100 data points for each class are only shown for graphical clarity) is shown in Figure 3.10 together with the decision boundaries, $x = \tau$, $\Psi_{Bayes}(x, qm)$ and $\Psi_{SVM}(x, qm)$. The data is clearly better separated in the evidence space using $\Psi_{Bayes}(x, qm)$ and $\Psi_{SVM}(x, qm)$ than using $x = \tau$.

Figure 3.9 shows the class-conditional marginal distributions of $p(x|A)$, $p(x|B)$, $p(qm|A)$ and $p(qm|B)$. Note that the marginal distributions of qm indeed do not provide any class discrimination. The classification results using all three decision boundaries are summarized in Table 3.1. The results are given in terms of class errors ER_A and ER_B, and total classification error ER. The results show that the total classification error obtained by applying both $Q - stack$ decision boundaries

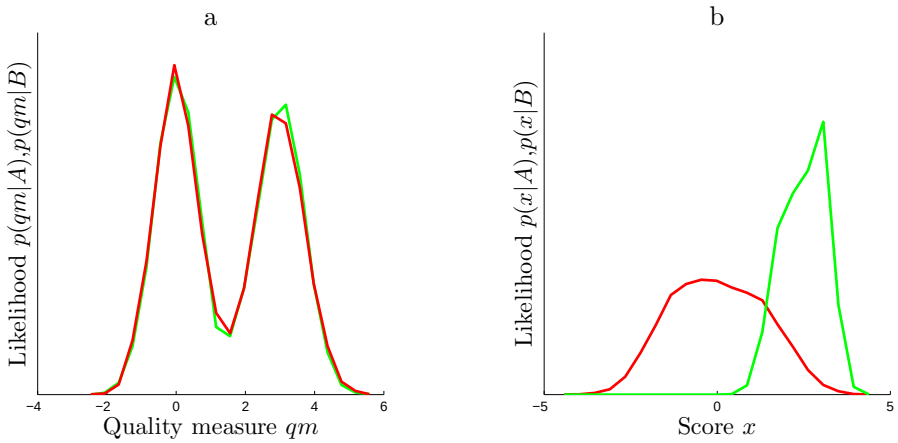


Figure 3.9: Example 1, linearly separable case, marginal evidence distributions.

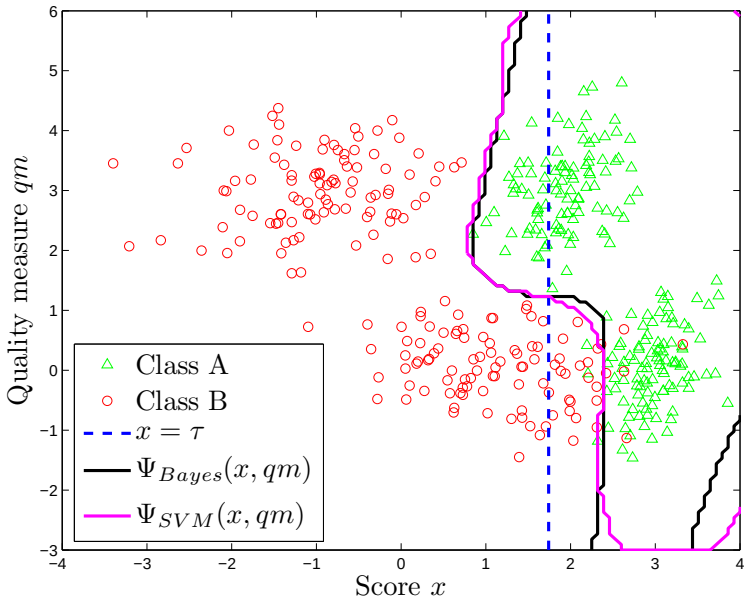


Figure 3.10: Example1: Classification in the evidence space $\mathbf{e} = [x, qm]$ for a linearly separable case.

$\Psi_{Bayes}(x, qm)$ and $\Psi_{SVM}(x, qm)$ is much smaller than the error of classification in the domain of x.

Table 3.1: Comparison of classification results in the score space $e = [x]$ and in the two-dimensional evidence space $\mathbf{e} = [x, qm]$, Example 1.

	ER_A	ER_B	ER
$x = \tau$	0.0059	0.2624	0.1342
Bayes/GMM	0.0167	0.0503	0.0335
SVM	0.0123	0.0544	0.0333

Example 2: Linearly nonseparable case

In this example we show how improved class separation and consequently classification can be achieved in the evidence space of $\mathbf{e} = [x, qm]$ for linearly nonseparable classes A and B. Consider data generated by a linear mixture of two Gaussian processes defined by parameters:

$$\Sigma_{A,1} = \begin{pmatrix} 0.5 & 0.095 \\ 0.095 & 0.2 \end{pmatrix}, \Sigma_{A,2} = \begin{pmatrix} 0.4 & 0.04 \\ 0.04 & 0.1 \end{pmatrix}, \mu_{\mathbf{A},\mathbf{1}} = \begin{pmatrix} 3 & 2 \end{pmatrix}^T, \mu_{\mathbf{A},\mathbf{2}} = \begin{pmatrix} 0 & 0 \end{pmatrix}^T.$$

$$\Sigma_{B,1} = \begin{pmatrix} 0.5 & 0.063 \\ 0.063 & 0.8 \end{pmatrix}, \Sigma_{B,2} = \begin{pmatrix} 0.4 & -0.18 \\ -0.18 & 0.9 \end{pmatrix}, \mu_{\mathbf{B},\mathbf{1}} = \begin{pmatrix} 3 & -1 \end{pmatrix}^T, \mu_{\mathbf{B},\mathbf{2}} = \begin{pmatrix} 0 & 3 \end{pmatrix}^T.$$

The data distribution (100 data points for each class are only shown for graphical clarity) is shown in Figure 3.12. Like in Example 1, the data is classified using three decision boundaries, $x = \tau$, $\Psi_{Bayes}(x, qm)$ and $\Psi_{SVM}(x, qm)$, and the three used decision boundaries are also shown Figure 3.12. It is clear that the two classes are not separable in the domain of x alone and that the addition of the dimension of qm is necessary for successful classification.

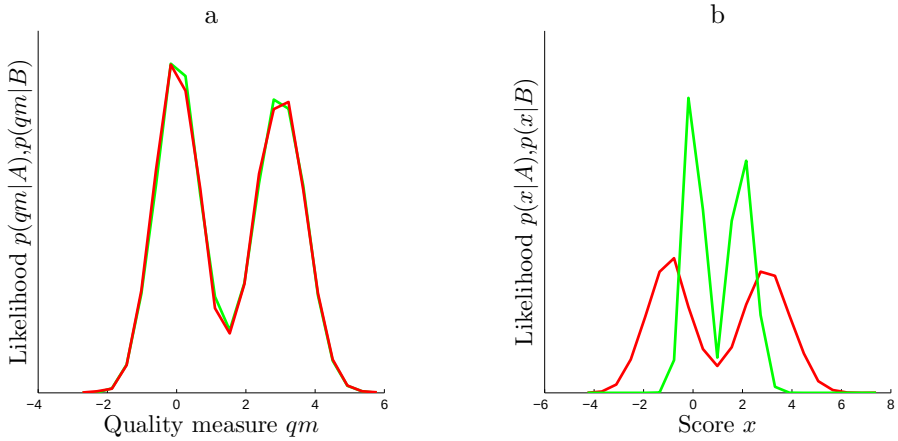


Figure 3.11: Example 2, linearly nonseparable case, evidence distributions.

Figure 3.11 shows the class-conditional marginal distributions of $p(x|A)$, $p(x|B)$, $p(qm|A)$ and $p(qm|B)$. The marginal distributions of qm do not allow for classification by themselves. The classification results using all three decision boundaries are summarized in Table 3.2. Note that the data separation in the score domain x was not feasible, which is also clear from the marginal

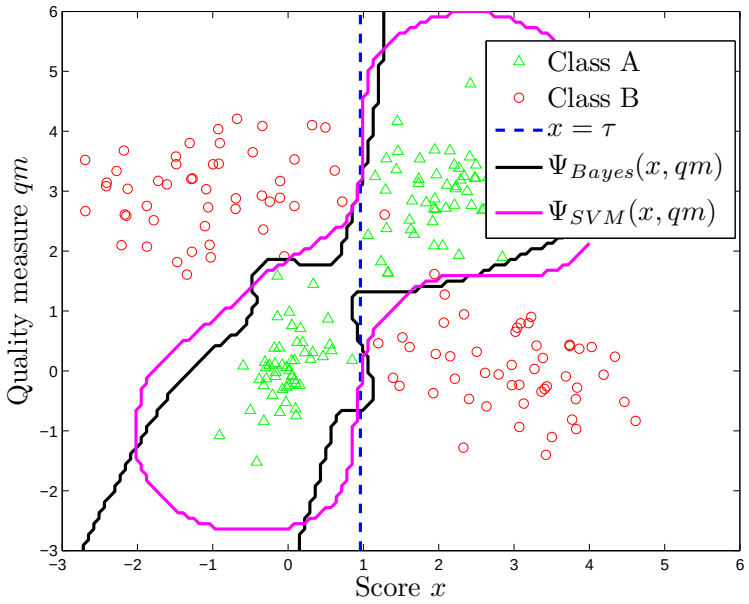


Figure 3.12: Example2: Classification in the evidence space $\mathbf{e} = [x, qm]$ for a linearly nonseparable case.

distributions shown in Figure 3.11. In the evidence space $e = [x, qm]$ the classification was possible despite the fact that the added dimension of qm does not provide class separation by itself. The tabelarized error rates confirm the converging conclusions drawn from the observations of the plots shown in Figure 3.12.

Table 3.2: Comparison of classification results in the score space $e = [x]$ and in the two-dimensional evidence space $\mathbf{e} = [x, qm]$, Example 2.

	ER_A	ER_B	ER
$x = \tau$	0.4983	0.5091	0.5037
Bayes/GMM	0.0121	0.0212	0.0167
SVM	0.0104	0.023	0.0167

3.5 $Q - stack$: improving classification with quality information

In Section 3.4 we have discussed how dependencies between features impact class separation and how augmenting the dimensionality of the classification space by a new, class-independent degree of freedom can afford better classification accuracy. In this section we will discuss how to practically employ these findings in a pattern classification system.

Note that in Section 3.4 all along we have been assuming that the available classification scores x and quality measures qm are simply random variables with certain distribution parameters. This is sufficient from the standpoint of the classification in the evidence space, but in a real application both

scores x and qm originate from certain stages of the signal processing involved in the classification process and as such they deserve a closer inspection here.

Consider the classification system presented in Figure 3.13. A biometric signal is registered and passed to the feature extraction module. From there, the extracted features are passed to the classifier unit which returns classification similarity scores x. In a conventional single-classifier system a final decision is reached by comparing x to a preset threshold τ. However, in the scenario proposed in this thesis the decision does not happen until a new classification step in the evidence space is taken. Namely, the scores x are further passed to an evidence combiner which builds an evidence vector out of the evidence x and class-independent quality measures qm. The evidence vector is then classified using a new classifier, whose yield is used to arrive at the final classification decision concerning the class label for the observed signal.

A classifier ensemble structure in which the output of one classifier is consequently used as a feature for the next is referred to as classifier stacking [122, 165]. The architecture proposed in this thesis bears a close resemblance to this idea. The major difference is that the stacked classifier's features are here augmented by class-independent quality measures qm. We hence refer to this new architecture as $Q-stack$: *classifier stacking with quality measures.*

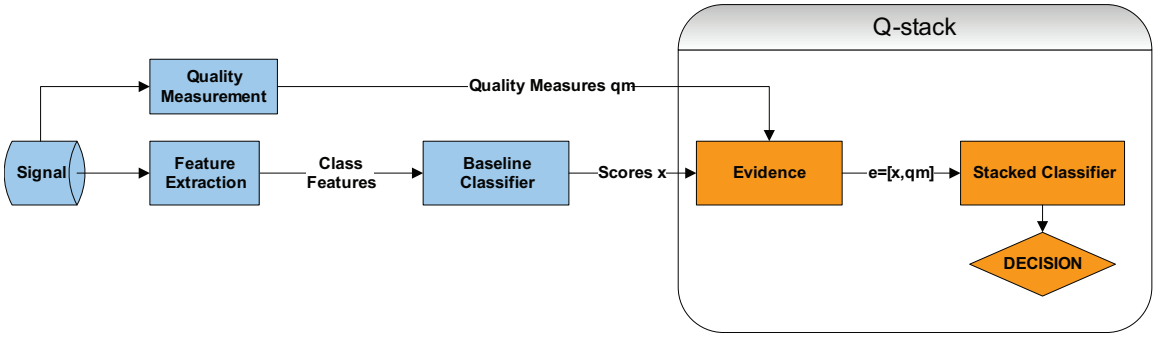


Figure 3.13: $Q-stack$ architecture for classification with class-independent quality information, for a single baseline classifier.

The $Q-stack$ architecture shown in Figure 3.13 is a practical embodiment of the theoretical findings presented in Section 3.4. Note that no assumptions whatsoever are being made as to the nature of the baseline classifier or the stacked classifier. The actual classifier type of the baseline classifier depends solely on the actual data classified and its choice is governed by general practices in the field of pattern classification [42, 111, 152]. The stacked classifier is chosen depending on the knowledge about the output of the base classifier and available quality measures. Both base and the stacked classifier may be constructed using statistical approaches, where the classifier parameters are learned from available training data, or using heuristic approaches, where the classifiers are constructed according to the prior knowledge of the classification problem. In the course of this thesis we shall focus on the trained, statistical approaches since they are typically better performing than concurrent heuristic approaches [43, 83].

3.6 $Q-stack$ architectures in single- and multiple classifier systems

As we have shown in Section 3.4.4 a straightforward extension from two to multiple dimensions of the evidence space will preserve the benefits that the intra-dimensional dependencies offer in

terms of class separation. As we have pointed out, each of the add-on dimensions to the evidence space must feature a dependency on at least one another dimension in order to be beneficial for the class separation. Figures 3.14 and 3.15 show structural extensions to the $Q-stack$ architecture presented in Figure 3.13 that account for the different possibilities of increasing the dimensionality of the evidence space.

Figure 3.14 shows an architecture of $Q-stack$ where one signal is classified by an ensemble of multiple classifiers (in the example shown: two classifiers). In an example involving a biometric verification scenario, the same face image could be classified by two different classifiers based on two disjoint sets of extracted features. Both classifiers return similarity scores x_1 and x_2, respectively. The quality of the signal qm is measured and passed to the evidence combining module. The evidence vector is constructed as $\mathbf{e} = [x_1, x_2, qm]$. According to the findings from Section 3.4.4, the quality measure must be dependent on either x_1 or x_2. By default x_1 is dependent on x_2 since they both originate from classifying the same signals. The amount of dependence between x_1 or x_2 is a measure of diversity between the baseline classifiers involved [96].

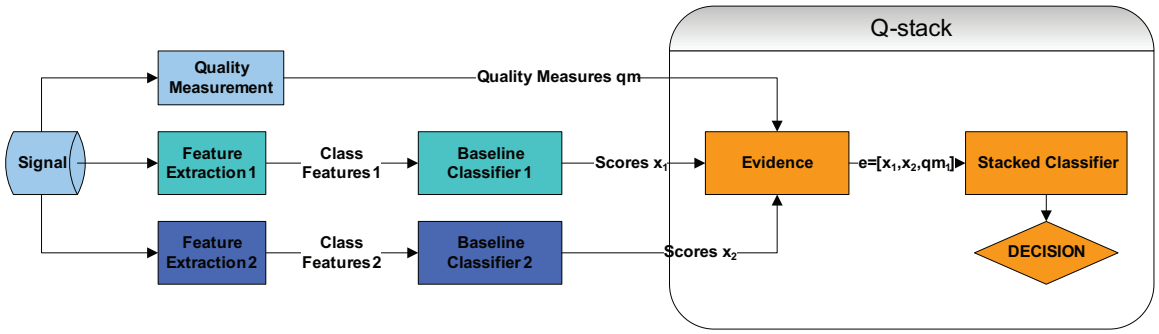


Figure 3.14: $Q-stack$ architecture for classification with class-independent quality information, for a multiple baseline classifier ensemble.

Figure 3.15 shows an architecture of $Q-stack$ for a multimodal scenario. Here, each modality is instantiated by two separate signals that pertain to one common physical phenomenon. In the biometric applications different biometric traits are typically regarded as separate modalities that pertain to the individuality of a person. For instance, fingerprint and face images are different signals yet are intrinsically linked to an individual from whom they were collected. As shown in Figure 3.15, both signals are processed and classified separately, producing scores x_1 and x_2. Two quality measures qm_1 and qm_2, one for each modality, is passed to the evidence combining module. Here a final evidence vector $\mathbf{e} = [x_1, x_2, qm_1, qm_2]$ is classified in order to arrive at the final classification decision. Again, here the evidence dimensions are pair-wise dependent, including a default dependence between x_1 and x_2 *.

It may at first appear strange to put together all available information, scores and quality measures, pertaining to different signals, into one evidence vector. However, at the time of building the evidence vector all available information is considered as orthogonal and relevant to the actual physical phenomenon (in biometrics: identity of a person). The ordering of the information in the vector is unimportant as long as the positions of each particular information in the vector is constant over all observations.

*The actual signals originating from each modality are likely to be independent since they represent different physical manifestations of the same phenomenon - here: identity of a person. However, the classification scores are bound to be dependent because they both reflect not the identity itself but the relationship with one of the two classes - imposter or client.

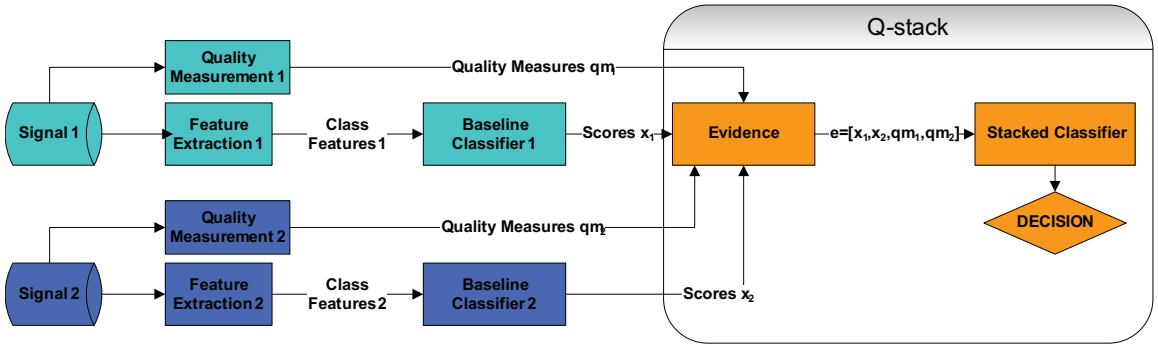


Figure 3.15: $Q-stack$ architecture for classification with class-independent quality information, for a multimodal classification system.

It is easy to construct further extensions to the $Q-stack$ scheme based on the architectures shown in Figures 3.14 and 3.15. For example, if more than one relevant quality measure is available for the given signal, each of them can be directly fed into the evidence combiner. Naturally, the new dimensions cannot be multiplied ad infinitum - it is a known fact that multiple correlated dimensions may create problems when constructing classifiers. We will return to this issue in Section 3.7.3.

In case when no quality information is available whatsoever, the evidence vector will contain only scores originating from individual classifiers. In this case the presented approach is equivalent to a trained classifier combination, which has been widely used in the field, for example for multimodal classifier fusion (also referred to as *score fusion*) [16].

We stress the fact that there is no conceptual difference between using $Q-stack$ in single- or multiple classifier systems, with or without quality measures.

3.7 $Q-stack$ as a generalized approach to classification with quality measures

As mentioned before in Chapter 2, several authors have proposed to use quality measures in order to improve classification performance. It is therefore important that we here discuss how the proposed method of $Q-stack$ relates to existing methods found in the subject literature.

In Chapter 2 we have shown a clear division between single-classifier systems and classifier fusion with quality measures. This division is an artificial one, as we have discussed in the Section 3.6. Nevertheless, here we wish to return to it as it reflects the currently established nomenclature in the field.

Existing literature seems to be making a clear distinction between multiple- and single classifier systems with quality measures, and the conceptual link between them seems to be ignored altogether. In previous Sections we argued that in fact there is no fundamental or conceptual difference between single-, multiple-, or multimodal classifier systems with quality measures. Here we will demonstrate it using examples of systems existing in the biometric literature.

3.7.1 Single classifier systems with quality measures

As we have discussed in Chapter 2, existing statistical methods of improving the performance of a single classifier with quality measures involve either a threshold selection [85, 163], or a model selection [127]. The selection is performed depending on the quality of the observed signal. Distribution

scaling [7] can be considered as a continuous linear model adaptation.

Consider the distributions shown in Figure 3.16. The figure presents three distributions of scores x drawn for signals registered at three different levels of quality, $qm_1 < qm_2 < qm_3$. For each of the three discrete levels of quality measure qm the class-conditional distributions of $p(x|A)$ and $p(x|B)$ change in shape and position in respect to origin of the horizontal axis of scores x.

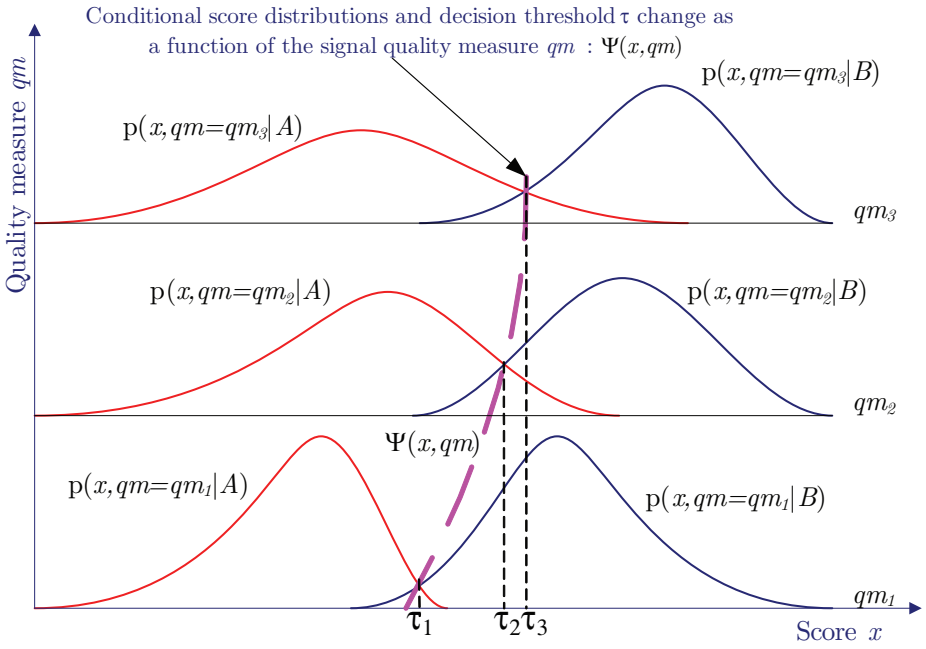


Figure 3.16: Quality-dependent threshold and model shift. Note that the models and respective decision thresholds drawn for discrete values of qm are samples that approximate the decision boundary $\Psi(x, qm)$ in the evidence space $e = [x, qm]$.

Assume now that for each of the discrete levels of quality measure qm_1, qm_2, qm_3 a set of two class-conditional models is created. In quality-dependent model selection procedures during the classification process a set of models is selected according to the observed discrete (or discretized) quality measure. A similar mechanism exists for the adaptive quality-dependent threshold selection. The decision thresholds for each quality level qm_1, qm_2, qm_3 are in fact thresholds estimated using certain models, again trained or otherwise built for each of the observed quality level. During the classification phase a decision threshold τ_1,τ_2 or τ_3 is selected depending on the observed quality measure for the new classified test sample.

If all quality-dependent models, or corresponding decision thresholds are placed in a two-dimensional coordinate systems then the resulting picture bears a striking resemblance to the evidence space $\mathbf{e} = [x, qm]$ discussed in Section 3.4. The main difference is that in model or threshold selection techniques the levels of quality are often discrete and not continuous.

In this light, the quality-dependent model and threshold selection techniques provide a sampled version of the decision boundary $\Psi(x, qm)$ drawn in the evidence space. This decision boundary is sampled at discrete points corresponding to the chosen discrete levels of quality measure qm. The finer the grid of discrete levels of qm the more finer and presumably accurate the approximation of $\Psi(x, qm)$ will be.

In most practically encountered signals there is no good reason to set a fixed set of discrete quality measures instead of permitting a continuum along the qm dimension. For example, acoustic

level of noise present in speech does not have a physically defined granularity that would suggest specific discrete levels of a corresponding quality measure. Here we arrive at one important drawback of the model and threshold selection schemes. Namely, in order to aptly and accurately estimate model parameters for a given discrete level of quality measure a certain minimal volume of training data is required. In many disciplines, including biometric authentication, the amount of available training data is frequently limited. In this case either one runs a risk of model overfitting, or he has to decide to cluster available datapoints whose quality measure lies in the vicinity of the chosen discrete qm. Such clustering has been adopted for instance in [65]. The clustering reduces the risk of overfitting at the cost of reduced resolution in the dimension of qm. The same or related problems haunt the quality-dependent distribution scaling methods [7]. In general, those methods attempt to approximate the decision boundary $\Psi(x, qm)$ in the evidence space by applying a linear transformation to the assumed baseline class-conditional distributions $p(x|A)$, $p(x|B)$ drawn for a reference signal quality level.

Since all methods mentioned above create an approximation of the optimal decision boundary $\Psi(x, qm)$, they may as well create an arbitrarily good approximation of it, contingent on the availability of data and appropriateness of assumptions made. In this sense there is nothing inherent to these methods that would render them mathematically inferior to the proposed method of $Q-stack$. Having said that, there are following major winning points for the technique proposed in this thesis:

- $Q-stack$ requires a creation of one classifier in a multidimensional space rather than creating multiple classifiers in one-dimensional space. This accounts to a greater simplicity and flexibility in classifier choice.

- $Q-stack$ can flexibly adopt multiple quality measures relevant to one observed score. It is not obvious how existing techniques of model/threshold selection could cope with that because a linear increase in the number of available quality measures would geometrically enlarge the amount of necessary thresholds or models.

- $Q-stack$ permits an easy evaluation of the potential performance of a given quality measure - classifier ensemble. For one single - classifier - single quality measure combination the gains from classification in the evidence space can be evaluated even without actually building a classifier: they are readily visible in the scatter plot of the available training evidence (for examples see Section 3.4).

3.7.2 Multiple classifier and multimodal systems with quality measures

Because of the fact that there is more than one classifier involved, the multi-classifier systems and multimodal systems traditionally were treated separately from single classifier systems. This division was inherited by the domain of biometric identity verification and existing literature seems to be making a clear distinction between them. In particular a term 'fusion' was coined for classifier combination. In particular, multimodal fusion has recently attracted a lot of attention in the domain of multimodal biometrics. Here, we continue the reasoning presented in Section 3.5 and show that the existing methods of multi-classifier combination with quality measures are special cases of the framework proposed in this thesis.

As an example let us consider a method of combining multiple classifiers with the use of quality measures proposed by Fierrez-Aguilar et al. to solve the problem of combining scores from two different fingerprint matchers [49]. The authors propose to use a combination function:

$$s_Q = \frac{Q}{2} s_M + (1 - \frac{Q}{2}) s_R, \tag{3.19}$$

where s_M and s_R are similarity scores normalized in the $s_M, s_R \in (0,1)$ range, originating from two different classifiers employed to compare the same pair of fingerprints. Q is a normalized quality measure for the given pair of compared fingerprints, $Q \in (0,1)$, and s_Q is the new, combined score. While the authors call the function given by Equation 3.19 an "*adaptive quality-based fusion strategy*", in fact it is an a-priori, heuristically defined decision surface in an evidence space made of $\mathbf{e} = [s_M, s_R, Q]$, as seen from the perspective of the framework of $Q - stack$. This decision hyperplane is shown in Figure 3.17.

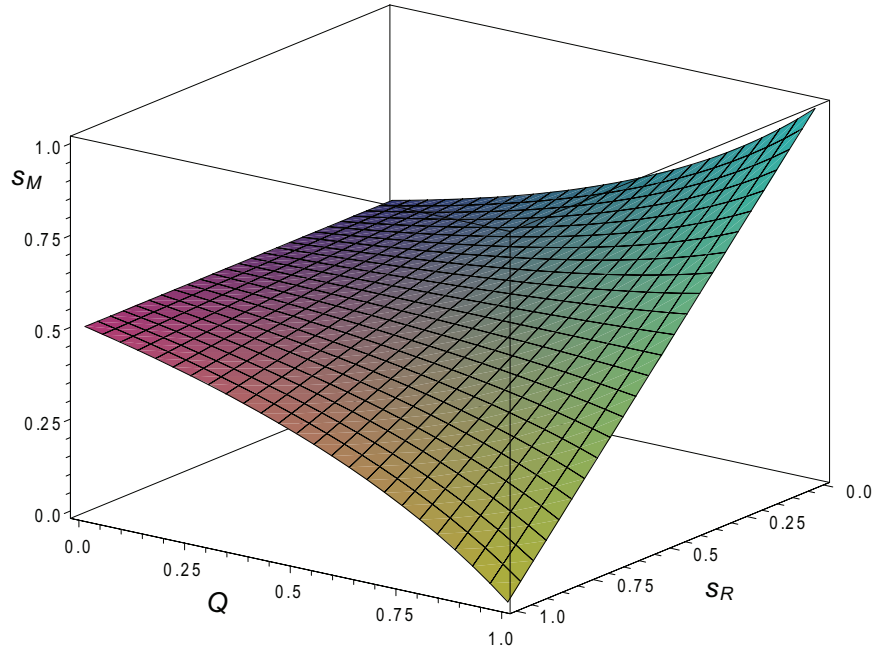


Figure 3.17: A graphical representation of the "*adaptive quality-based fusion strategy*" from [49], which is in fact a heuristically defined decision boundary in the evidence space $e = [s_M, s_R, Q]$.

Note that despite the fact that the authors do not make any explicit claims as to the statistical dependencies between Q, s_M and s_R, the definition of the decision boundary tacitly assumes a linear correlation between s_M and s_R as it can be easily derived from Equation 3.19, and is readily visible in Figure 3.17.

The example is therefore a special case of classification in an evidence space defined by the dimensions of two classifier scores and a class-independent quality measure, discussed in detail in Section 3.4.4. Since the method proposed in [49] is heuristic in nature, it defines a priori the shape of the decision boundary while disregarding the actual shapes of the joint distributions of evidence. Other heuristically defined decision hyperplanes, like the polynomial fusion functions from [153] or approaches postulated by Fierrez-Aguilar in [48], can be represented in a similar way.

It therefore becomes clear that different methods of incorporating quality measures in the classification process, both in a single- or multiple-classifier scenario, are in fact different algorithmic realizations of the very same mathematical principles, given in Section 3.4.1. The conclusion follows that the problems of data dimensionality and of choosing a proper set of quality measures further discussed in this chapter are as valid for the generalized $Q - stack$ method presented in this thesis, as it is for other mentioned methods in the subject literature.

3.7.3 Problem of data dimensionality

Adding new score or quality measure features into the evidence vector inadvertently leads to an increased dimensionality of the classification problem. As a consequence of the so-called *dimensionality curse* [14, 42], in comparison with a certain baseline number of dimensions, increased problem dimensionality calls for exponentially increased volume of available data samples in order to achieve the same spatial sampling frequency. Adding new components to the evidence vector by default preserves the amount of data available for training the stacked classifier. As a consequence, increased problem dimensionality may lead to overtraining (model over-fitting) of the stacked classifier.

A typical solution for avoiding excessive data dimensionality is feature selection and dimensionality reduction. Classical dimensionality reduction techniques based on Principal Component Analysis (PCA) [42] search for the most compact data representation rather than increased class separation. Successful application of these techniques in order to reduce the dimensionality of the evidence vector depends solely on the actual joint distributions of $p(\mathbf{e}|A)$ and $p(\mathbf{e}|B)$. Similarly, any of the feature selection techniques based on the Linear Discriminant Analysis may or may not be a good choice. In general, if classes are linearly separable (like in Figure 3.10) in the evidence space then dimensionality reduction techniques based on linear projections can be expected to perform well. If classes cannot be well separated by a linear decision hyperplane in the evidence space (see example in Figure 3.12) then linear projection-based dimensionality reduction techniques are also bound to fail.

Dimensionality reduction techniques that operate in the evidence space take into account the relationships between the components of the evidence vector. However, there exists a family of feature selection techniques that consider the features individually, in detachment from the others [152]. These techniques inspect how much of a separation between classes can be achieved by each feature individually, and are equivalent to making the feature selection based on the marginal distributions alone. Since the quality measures are by definition class-independent, they will be swiftly eliminated from the feature pool. Therefore all techniques that perform feature selection based by looking at them individually are inadequate and inappropriate for the problem that is the topic of this thesis.

Quality measures are class-independent, but they are by design dependent on the classification errors of the baseline classifier, which is a consequence of their dependency on the classification scores x. For this reason in order to inspect how useful certain quality measures are individually, the way to do it is to inspect the class-conditional distributions of $p(qm|C=1)$ and $p(qm|C=0)$ where C is a binary random variable that takes the value of $C=1$ if the baseline classifier made a correct decision, and the value of $C=0$ if the baseline classifier made an incorrect decision. The estimation of $p(qm|C=1)$ and $p(qm|C=0)$ may be biased depending on the bias of the baseline classifier. Note that in the $Q-stack$ scheme the baseline classifier scores are not binarized by an application of a baseline threshold *. We will return to this topic in greater detail in Chapter 4.

The problem of excessive data dimensionality usually plays a role when the number of dimensions of the classification problem becomes large and comparable to the number of datapoints available for classifier training. In biometrics, which is the main focus of this work, one is unlikely to ever encounter such a vast number of concurrent quality measures and classifiers that feature selection techniques would have to be applied to the evidence vector. A possible solution to this problem has been proposed by Poh et al. in [121], where the quality measures are clustered into a form of discrete

*Although such an eventuality is conceivable, in such a case the binarized scores (decisions) would be combined with quality measures using a dedicated combining algorithm. Since it is hard to envision the benefits of doing so and score-based methods of classifier combination routinely perform better than decision based methods [91], we do not inspect this possibility in greater detail.

meta-quality states. We do not pursue the topic of feature selection any deeper in this thesis.

3.8 A closer look at noise and quality measures

So far we have been showing examples where quality measures qm were given a priori for every available score x. This is naturally not the case in real situations, where the quality measure needs to be estimated based on some kind of noise measurement. The a priori assumption of quality measures skips altogether two important issues:

1. how does the noise n actually interact with the scores x, and

2. how does an uncertainty in estimating qm impact classification.

Answering these questions is expected to shed light on a problem if using irrelevant or uncertain quality measures can actually hurt the classification performance. In this section we provide more insight into these problems. Again, we will use synthetic data examples in order to be able to freely manipulate the parameters of the underlying data- and noise generating processes. Real data does not offer such a comfort of experimentation - when using real data it is often not obvious if the obtained improvement over existing methods is systematic and generalizable. By using synthetic data we are able to draw theoretically sound generalizations.

In the examples given in Sections 3.4.7 and 3.4.7 we have been defining the covariance matrices of the joint class-conditional distributions of evidence $p(\mathbf{e}|A)$ and $p(\mathbf{e}|B)$. In this section we shall neither define a priori the correlations between the component of the evidence vector, nor generate the quality measures themselves. Instead, we assume only the parameters of the stochastic processes A and B that generate noiseless scores given by $p(x'|A)$ and $p(x'|B)$, and the parameters of a noise-generating process N that generates class-independent noise instances n. The actual observed scores x are the yield of a function $x = \Phi(x', n)$. In the experiments reported in this section we will use an additive model of noise, $\Phi(x', n) = x' + n$ since it is simple enough to analyze algebraically and is a type of noise frequently observed in practice, for instance in speech recognition [56]. The choice of the nature of function Φ is immaterial from the viewpoint of presented methodology of $Q - stack$ and its notion is never used throughout the experiments presented here. Instead, as it was discussed earlier in this chapter, the relationship between X' and N is implicitly learned from available data. In practice, some knowledge of the nature of function $\Phi(x', n)$ is prerequisite for all heuristic methods discussed earlier in Section 3.7.

The exact value of n would make a perfect quality measure qm. However, it would be rather idealistic to assume that noise could be measured without any level of uncertainty or error. Therefore here we will examine how imprecisions in noise estimation impact the classification performance. In the experiments presented here the observed quality measures $qm = n + d$ are uncertain measurements of the noise instances n, where d is an instance of a random process D. We will assume d to be normally distributed. The process D therefore represents the uncertainty in estimating qm.

The data flow diagram of the experiments presented in this section is shown in Figure 3.18. The parameters of processes A, B, N and D as well as the nature of the function $\Phi(x', n)$ are used exclusively for the purpose of data generation and are never used in order to adjust the parameters of classifiers applied.

We use four different classifier types as stacked classifiers:

1. Linear Discriminant Analysis - based classifier: LDA

2. Quadratic Discriminant Analysis - based classifier: QDA

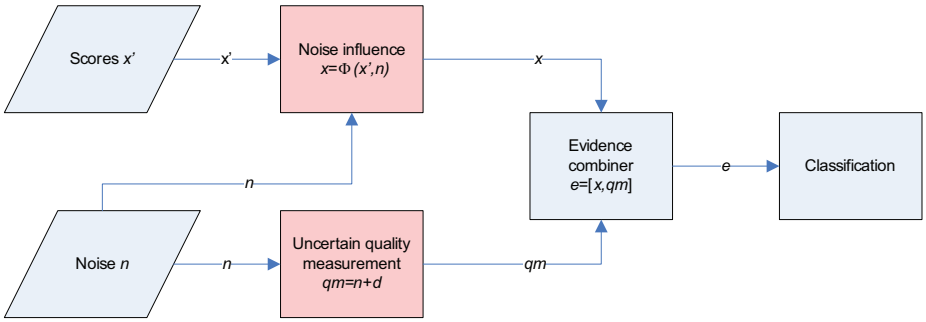


Figure 3.18: Diagram of the data flow of the experiment, Section 3.8.

3. Bayes classifier using Gaussian Mixture Model - based distribution representation: Bayes

4. Support Vector Machines - based classifier using RBF kernel: SVM

The classifiers are trained using 1000 training data points and then deployed to classify another 1000 testing data points. Having the knowledge of the underlying data-generating stochastic processes at hand it is easy to compute that the optimal decision function that separates $p(\mathbf{e}|A)$ and $p(\mathbf{e}|B)$ is linear. However, we still insist on using not only the LDA classifier, which inherently generates linear decision boundary. Other classifiers are also employed in accord with the previous premise that the notion of the nature of $p(\mathbf{e}|A)$ and $p(\mathbf{e}|B)$ should not be used heuristically in the classification process. For the same reason, despite the fact that the underlying distributions are Gaussian we forced a four-component GMMs as distribution models used by the Bayes classifier, which led to a certain sub-optimality of its performance.

Consider following Gaussian processes:

$$\begin{aligned} p(x'|A) &= \mathcal{N}(\mu_{x',A}, \sigma^2_{x',A}), \mu_{x',A} = -1, \sigma^2_{x',A} = 1, \\ p(x'|B) &= \mathcal{N}(\mu_{x',B}, \sigma^2_{x',B}), \mu_{x',B} = 1, \sigma^2_{x',B} = 1 \end{aligned} \tag{3.20}$$

The Bayes error associated with the classification of x' into classes A and B can be analytically estimated and is:

$$E'_{Bayes} = \frac{1}{2}\int_{-\infty}^{\tau} p(x'|B)dx' + \frac{1}{2}\int_{\tau}^{\infty} p(x'|A)dx' \approx 0.1587, \tag{3.21}$$

where $\tau = 0$ is the decision threshold in the x' domain. Let the noise-generating process N produce noise instances n according to

$$p(n) = \mathcal{N}(\mu_N, \sigma^2_N), \mu_N = 0, \sigma^2_N = 1. \tag{3.22}$$

If no noise would be present, observed scores would be $x = x'$. Assume that in the presence of noise N the observed scores x are affected by the noise n according to $x = \Phi(x', n) = x' + n$. consequently the class-conditional distributions of observed scores $p(x|A)$ and $p(x|B)$ are given by convolution of the probability density functions [63]:

$$\begin{aligned} p(x|A) = p(x' + n|A) = p(x'|A) * p(n|A) = \mathcal{N}(\mu_N + \mu_{x',A}, \sigma^2_N + \sigma_{x',A}) \\ p(x|B) = p(x' + n|B) = p(x'|B) * p(n|B) = \mathcal{N}(\mu_N + \mu_{x',B}, \sigma^2_N + \sigma_{x',M}) \end{aligned} \tag{3.23}$$

The associated Bayes error is then

$$E_{Bayes} = \frac{1}{2}\int_{-\infty}^{\tau} p(x|B)dx + \frac{1}{2}\int_{\tau}^{\infty} p(x|A)dx \approx 0.2398, \tag{3.24}$$

for the decision threshold again equal to $\tau = 0$, since $\mu_N = 0$ does not shift the means of the distributions and consequently does not bias the classifier. Bayes error E_{Bayes} is the best class separation theoretically attainable when classifying the observed scores x alone.

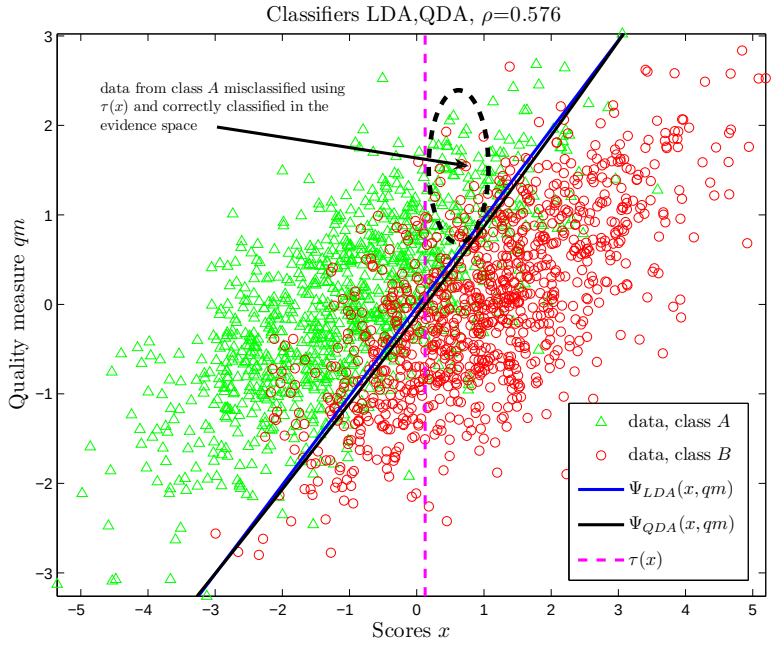
Let us now measure the quality measure qm. Ideally, we wish to measure the actual noise n directly. In this case $qm \propto n$. This ideal measurement may in practice be not feasible and the noise measurement may itself have some uncertainty to it. We model this possible uncertainty by adding white Gaussian noise of controlled variance σ_d^2 to the measurement of qm. In this scenario, for $\sigma_d^2 = 0 \Rightarrow qm \propto n$, and for $\sigma_d^2 \to \infty$ the quality measure qm becomes independent on the actual noise n, and thus it ceases to be informative from the viewpoint of classification using $Q - stack$. Since all involved processes are Gaussian then the dependency between quality measurements and scores can be measured by computing the correlation coefficient ρ between qm and x.

In the experiments shown in this section we classify 1000 testing data points, using classifiers trained on a separately generated set of 1000 training data points. The data are generated by processes described above. The impact of ρ on the class-conditional distributions evidence, $p(\mathbf{e}|A)$, $p(\mathbf{e}|B)$ is shown in Figures 3.19 and 3.20.

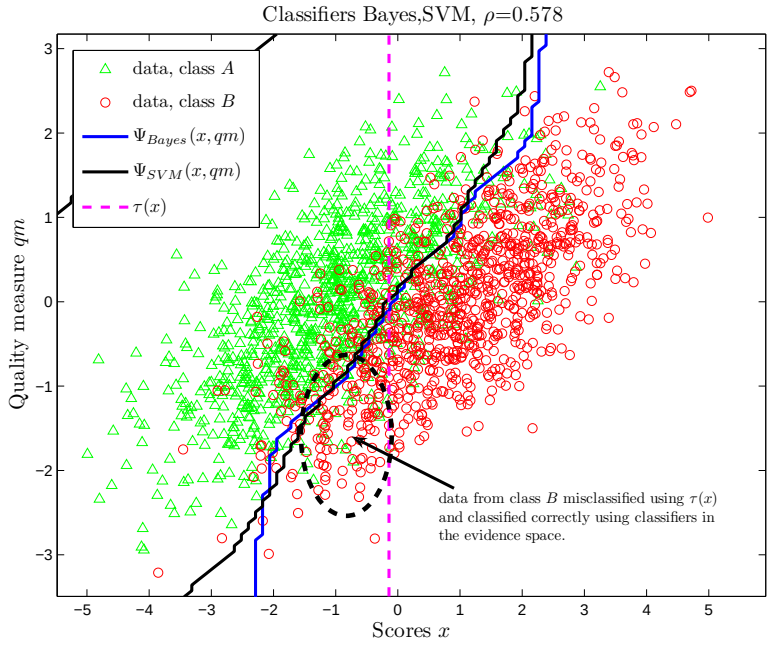
In Figure 3.19 the quality measures qm are not uncertain and they represent the idealistic case when the actual noise instances can be measured without error, $\sigma_d^2 = 0$. In the experiments shown in Figure 3.19 this resulted in the correlation coefficient between scores x and quality measures qm of $\rho \approx 0.58$. In order to best visualize the difference between the classification in the evidence space $\mathbf{e} = [x, qm]$ as opposed to $e = [x]$ corresponding decision boundaries are shown. The decision boundary corresponding to classification in the score domain at the decision threshold $\tau(x)$ is equivalent to the decision boundary $x = \tau$ in the evidence space $\mathbf{e} = [x, qm]$.

For clarity purposes, the $Q - stack$ decision bounds for four classifiers operating in the evidence space are presented in two separate subplots. Decision boundaries computed by the LDA and QDA classifiers are shown in Figure 3.19(a) and those computed by the Bayesian and SVM classifiers are shown in Figure 3.19(b). Note the data points whose approximate location is marked by dashed line ellipses - these data points are bound to be misclassified in the score domain by applying threshold τ. Thanks to the addition of the new dimension of quality measures qm these data points are classified correctly using the $Q - stack$ approach.

Figure 3.20 demonstrates graphically an example of the impact of the dependency between scores x and quality measures qm and of the uncertainty in estimating qm, on the classification in the evidence space $\mathbf{e} = [x, qm]$. In this Figure, the measurement of quality measure is very noisy at $\sigma_d^2 = 50$, resulting in a low correlation coefficient between x and qm of $\rho \approx 0.13$. Since $\mathbf{e}$ is normally distributed then low value of the correlation coefficient ρ necessarily entails low dependency between x and qm. This is reflected in the fact that the decision boundaries Ψ between classes A and B tend towards $x = \tau$, the decision boundary obtained when considering only x for classification. As the difference between classification in the evidence spaces of $\mathbf{e} = [x, qm]$ and $e = [x]$ wanes with growing σ_d^2, so does the benefit of using quality measure as an add-on dimension in the evidence vector. With neither inherent class-selective information nor dependency on scores x the extra dimension becomes not useful. At the same time though, addition of an irrelevant quality measure is not causing a degradation of class separation in the evidence space in comparison with using scores x alone.

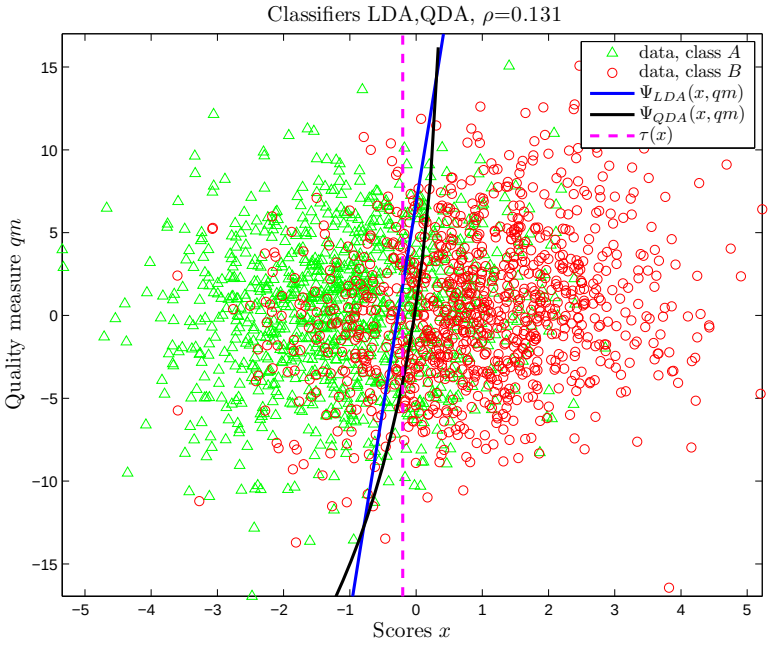


(a) Classifiers $\tau(x)$, LDA and QDA

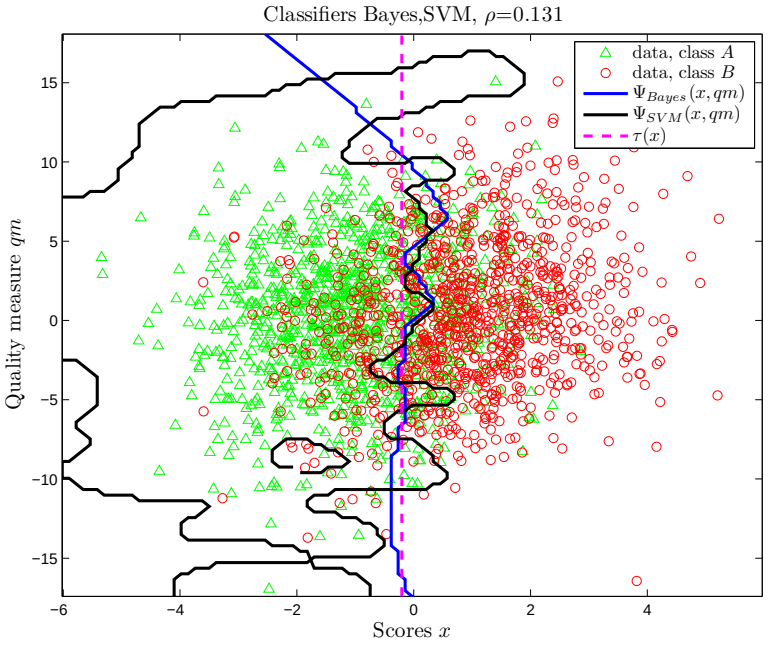


(b) Classifiers $\tau(x)$, Bayes and SVM

Figure 3.19: Class-conditional evidence distributions $p(\mathbf{e}|A)$ and $p(\mathbf{e}|B)$ with $Q - stack$decision boundaries for LDA, QDA, SVM and Bayes classifiers. Quality measures taken at $\sigma_d^2 = 0$.



(a) Classifiers $\tau(x)$, LDA and QDA



(b) Classifiers $\tau(x)$, Bayes and SVM

Figure 3.20: Class-conditional evidence distributions $p(\mathbf{e}|A)$ and $p(\mathbf{e}|B)$ with $Q - stack$decision boundaries for LDA, QDA, SVM and Bayes classifiers. Quality measures taken at $\sigma_d^2 = 50$

Compare the behavior of the decision boundary Ψ_{SVM} in Figures 3.19(b) and 3.20(b). The curve shown in Figure 3.20(b) shows a clear overfitting to the training data as a result of an increase in dimensionality of $\mathbf{e}$ beyond necessity. This is not the case in figure 3.19(b). Such overfitting may be avoided by quality measure clustering tricks, as shown in [121], or by simply choosing a classifier of a smaller parametric complexity.

Figure 3.21 presents the explicit relationship between the correlation coefficient ρ between x and qm and the classification error rates in the evidence space using decision boundaries $\tau(x)$, Ψ_{LDA}, Ψ_{QDA}, Ψ_{Bayes} and Ψ_{SVM}. The parameters of A, B and N were kept constant over the entire experiment. The variance σ_d^2 of the process D that adds uncertainty to the measurement of qm was changed in the range (0:20). Figure 3.21(a) shows the classification errors in the terms of mean HTER after 50 independent experimental runs, for the training set, and Figure 3.21(b) shows the classification errors for the testing set, also in terms of mean HTER. The error bars show the standard deviation of obtained corresponding error values after 50 experimental runs. The numerical results of this experiment are gathered in Table 3.3. Note that for compactness we decided to give the numerical results only for a selected representative subset of σ_d^2 values shown in Figure 3.21. In Table 3.3 the reader will find included classification results using only quality measures qm. These results have been included here in order to show that indeed the quality measures qm by themselves do not allow for class separation - note that the $HTER \approx 0.5$ for classification in the $e = [qm]$ domain regardless of the value of ρ.

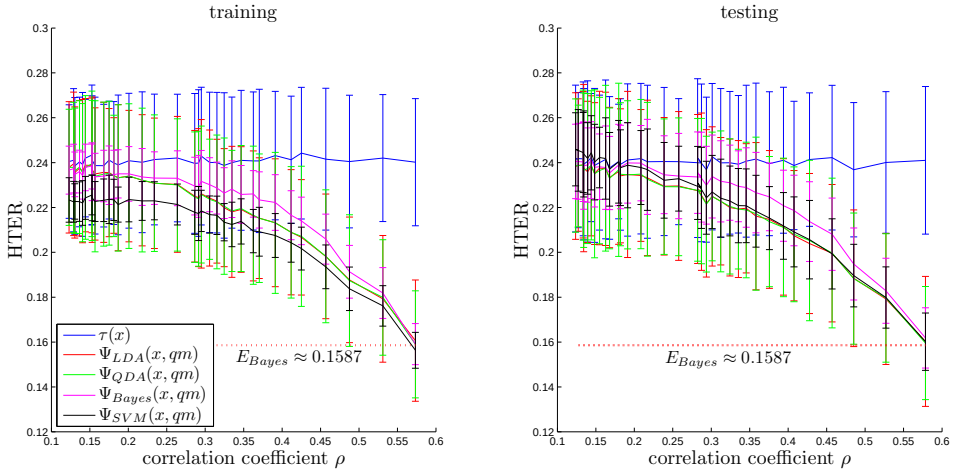


Figure 3.21: Impact of the correlation ρ between the observed scores x and the observed quality measures qm , for additive noise.

The classification results obtained are bounded by the Bayes errors given by Equations 3.21 and 3.24. The classifier that operates only in the score domain, $\tau(x)$, achieves a classification performance comparable with the theoretically computed value of E_{Bayes} given by Equation 3.24. All classifiers that operate in the evidence space obtained classification results comparable to the theoretically computed value of E'_{Bayes} given by Equation 3.21 for $\sigma_d^2 = 0$. In other words, the $Q - stack$ method allowed for a recovery of the Bayes error E'_{Bayes} associated with the noiseless data although noisy data was used for both classifier testing and for training. With the growing value of σ_d^2 the quality measures qm became increasingly less informative to reach obtained classification $HTER$ in the vicinity of E_{Bayes}.

σ_d^2	0	0.4	1	2.6	7	15	20
$\overline{\rho(n,qm)}$	1	0.845	0.7069	0.5259	0.355	0.2513	0.2184
$\overline{\rho(x,qm)}$	0.5785	0.4852	0.4075	0.3012	0.2086	0.1441	0.1272
$HTER$							
$\mu_{\tau(qm)}$	0.4978	0.4995	0.499	0.5003	0.5002	0.5019	0.5007
$\sigma_{\tau(qm)}$	0.0354	0.0334	0.0335	0.0449	0.0326	0.036	0.0383
$\mu_{\tau(x)}$	0.241	0.2368	0.2387	0.2426	0.2418	0.2423	0.2411
$\sigma_{\tau(x)}$	0.0329	0.0299	0.0286	0.0326	0.0336	0.0286	0.0302
μ_{LDA}	0.1603	0.1884	0.2074	0.2249	0.2343	0.2386	0.2389
σ_{LDA}	0.029	0.0305	0.0291	0.031	0.0294	0.0312	0.0295
μ_{QDA}	0.1596	0.1883	0.208	0.2249	0.2349	0.2388	0.2385
σ_{QDA}	0.0252	0.0293	0.0301	0.0287	0.0333	0.0333	0.0271
μ_{Bayes}	0.1618	0.1948	0.2186	0.2338	0.2398	0.2411	0.2406
σ_{Bayes}	0.0134	0.016	0.0157	0.0185	0.0168	0.0171	0.0169
μ_{SVM}	0.1602	0.1896	0.2091	0.2269	0.2378	0.2444	0.2455
σ_{SVM}	0.0128	0.0139	0.0156	0.0163	0.0165	0.0165	0.0182

Table 3.3: Selected $HTER$ results from Figure 3.21(b), 1000 data points, mean values and standard deviations after 50 repetitions for each value of σ_d^2.

The results shown in Figure 3.21 and in Table 3.3 confirm the theoretical findings given in this Chapter. In this section we have given an example using an additive noise model in order to exemplify our theoretical findings. While we have not used the information about the noise nature in constructing the example presented here, the reader might be curious about how other noise types could be handled by the presented approach. For this reason we are giving an example using multiplicative noise model in Appendix B. As the reader will see further in this thesis, both additive and multiplicative noise models result in class-conditional evidence distributions of synthetic data that bear resemblance to the distributions observed in real data. Naturally it is not feasible to present here, neither mathematically nor using graphical representations, all possible types of distributions of evidence. However, the principles and algorithmic steps that govern classification using the $Q-stack$ approach can be applied and put to a practical use without the need of explicit mathematical notion of underlying functions or processes. In order to emphasize this, we devote Section 3.9 to the practical, algorithmic aspect of the proposed technique of $Q-stack$.

3.9 $Q-stack$ - a generalized algorithm

The results of experiments with synthetic data presented in Section 3.8 and in Appendix B confirm the theoretical predictions of Section 3.4.1. The use of quality measures that are statistically dependent on the similarity scores obtained from the baseline classifier allowed for classification in the evidence space with greater accuracy than in the score space alone. In the example given in Section 3.8 the accuracy gain of the classifier could be computed analytically since the all data-generating stochastic processes were also defined a priori and known. The dependency of qm on x could be also easily expressed in terms of correlation coefficient thanks to the fact that the data-generating processes were Gaussian.

In real-life application such a comfort is not readily available. Typically one has no precise notion

of neither the actual analytical descriptions of the joint distributions of neither $p(\mathbf{e}|A)$, $p(\mathbf{e}|B)$ nor $p(n)$. The uncertainty of estimating the value of noise n manifesting itself in the accuracy of qm is usually non-zero and the dependency of qm on x may be far more complex than a linear correlation. Even if that information was available, it is easy to imagine how the complexity of arithmetical representations could render any analytical derivations impractical.

However, as we have mentioned previously, the strength of the proposed approach is in the fact that its practical application does not require specific knowledge of the mathematical representations of statistical processes that govern the generation of the observed data. Instead, necessary dependencies can be implicitly encoded into the parametrization of the classification problem.

The key point of the proposed scheme of $Q-stack$ is classification of evidence vectors in the evidence space using a stacked classifier. Consider only this classifier. Upon having observed sufficient amount of training data, the stacked classifier can be constructed using any data-driven, statistical classification methods known in the domain of pattern classification [42, 111]. The choice of the classifier and its parameters is an implicit form of expressing all the arbitrarily complex dependencies between evidence space dimensions that prohibit a practical symbolic analysis. Instead, the classifier is numerically optimized for the available training data.

Algorithm of $Q-stack$

The pragmatic approach to the use of $Q-stack$ for classification can be broken down into a sequence of algorithmic steps that allow for a realization of the $Q-stack$ architectures given in Figures 3.13, 3.14 and 3.15 from Section 3.5. The algorithm consists of following steps:

Training

1. Classify available training data using the previously available k baseline classifiers in order to obtain scores $x_1^{TR}, ..., x_k^{TR}$.

2. For available training data, collect l quality measures $qm_1^{TR}, ..., qm_l^{TR}$ that are relevant (dependent) to the baseline classifiers.

3. Assemble training evidence vectors $\mathbf{e^{TR}} = [x_1^{TR}, ..., x_k^{TR}, qm_1^{TR}, ..., qm_l^{TR}]$ out of the observed baseline classifier scores x^{TR} and quality measures qm^{TR}.

4. Using the training evidence $\mathbf{e^{TR}}$, create a classifier that separates $\mathbf{e^{TR}}|A$ from $\mathbf{e^{TR}}|B$. The nature of the stacked classifier must be chosen according to the observed joint class-conditional evidence distributions $p(\mathbf{e^{TR}}|A)$ and $p(\mathbf{e^{TR}}|B)$.

5. If available, optimize the parameters of the stacked classifier on a separate development set, for improved generalization and in order to avoid model over-fitting.

Testing

1. Classify newly observed (tested) data using the same k baseline classifiers in order to obtain the scores $x_1^{TS}, ..., x_k^{TS}$.

2. Collect the relevant quality measures $qm_1^{TS}, ..., qm_l^{TS}$.

3. Assemble testing evidence vectors $\mathbf{e^{TS}} = [x_1^{TS}, ..., x_k^{TS}, qm_1^{TS}, ..., qm_l^{TS}]$ out of the observed baseline classifier scores x^{TS} and quality measures qm^{TS}.

4. Classify the testing evidence $\mathbf{e^{TS}}$ using the stacked classifier.

3.10 Summary

In this chapter we have presented the theoretical foundations of classification with class-independent quality information. We have explained the mechanism behind the observed occurrence of misclassifications due to the presence of noise. Departing from here, we have introduced the concept of statistical dependencies between the noise and the observed classification scores. We have described the need of collecting quality measures in order to quantify the impact of noise on the observed signals, and we have shown how the quality measures inherit dependencies on classification scores. We have introduced the concept of evidence as a vector of all available classification scores and quality measures pertaining to a given class observation.

Further we have addressed the problem of the impact of feature dependencies on class separation, with a particular focus on the use of class-independent information as one of the dimensions of the classification space. We have explained how an addition of class-independent information can afford improved class separation due to the existence of statistical dependencies between the quality measures and classifier scores. We discussed how quality measures, which by themselves are individually irrelevant from the viewpoint of the classification process, become conditionally relevant in the context of another features contained in the evidence vector. We have presented an analytical analysis of the impact of correlation between class-selective and class-independent dimensions in the classification space on class separation, using examples of bivariate normal distributions, and we have provided an intuition behind the observed results. Then, we have extended the conclusions of the theoretical findings from bivariate-normal to arbitrary evidence distributions.

Acknowledging the fact that deriving analytical solutions for arbitrary distributions is not always feasible, we have introduced a data-driven approach to classification with class-independent quality measures. The proposed approach is based on the concept of classifier stacking and involves a two-layer classifier ensemble. The first layer contains baseline classifiers and quality estimation modules, while the second layer consists of an evidence combiner and a stacked classifier. The nature of the stacked classifier is chosen depending on the class-conditional distributions of available training data. The proposed architecture, $Q - stack$, has been presented as a general framework of classification with quality measures. We have discussed how $Q - stack$ encompasses existing approaches towards classification with quality measures found in the prior art. We have shown how the same principles of $Q - stack$ apply to single-, multiple-classifier and multimodal systems alike.

Finally, we have analyzed the impact of uncertain quality measurements on the performance of the $Q - stack$ algorithm. Using synthetic data, we have simulated a range of uncertainty degrees in quality assessment and we have shown that as the quality measurements become increasingly random in respect to the actual noise present, the class separation and corresponding classification accuracy approach that of the baseline classifiers. Therefore we have shown that the use of uncertain or irrelevant quality measures does not hurt class separation, given an appropriate choice of the stacked classifier. At the same time, accurate quality measures have been demonstrated to recover the class separation of noise-free data without having previously observed it in the training phase.

Credence and single error prediction

4

4.1 Introduction

Classifiers and classifier ensembles are constructed with the aim to maximize classification accuracy. However, certain classification problems are inherently difficult and errors do occur. This is the case, for instance, in biometric identity verification, where the classification systems often have to contend with adverse environmental conditions. If errors are inevitable, it is important to ascertain the chances that any particular classification decision taken is correct. In the area of biometric identity verification a decision of accepting a particular identity claim may entail severe legal, financial, or personal consequences. Therefore in this field, the ability to accurately predict single classification errors is of paramount importance.

In Chapter 2 we have discussed the existing approaches towards error prediction and estimation. We have shown that only a handful of methods actually focus on estimating the probability of a single erratic classification decision. Out of the existing methods, few offer a sound probabilistic interpretation of the derived predictions, and none allows for an inclusion of signal quality information in the estimation process. As we have elaborated in Chapter 3, quality measures as estimators of adverse environmental conditions are closely linked with classifier errors in situations where the classified observations may be noise-contaminated. As such, it is clear that a reliable method of error prediction must be able to take into account the quality information.

In Chapter 3 we have developed a framework of classification with class-independent quality information, where quality measures are seamlessly incorporated in the classification process. The proposed framework, $Q - stack$, utilizes a stacked classifier ensemble in which the second-level, stacked classifier can be of arbitrary nature. If a Bayesian classifier is chosen as the stacked classifier, then the yield of this classifier is by default probabilistic, and can be interpreted as a *degree of belief* or *credence* in the correctness of choosing one class over another. As we have elicited in Chapter 2, the sought estimate of probability of single decision correctness is a subjective Bayesian credence.

In this chapter we develop these ideas and we present the link between the credence in choosing one class over another with the credence in decision correctness, which is closely linked with estimating the probability of a single classification error. We show that indeed a Bayesian classifier

employed as a stacked classifier in the $Q - stack$ framework produces posterior probabilities which can be successfully employed as single misclassification predictors. Since the proposed error prediction method is a natural extension of the $Q - stack$ framework, it is able to seamlessly include signal quality measures. Further, we propose evaluation criteria for the credence estimates and demonstrate the performance of the proposed method using a synthetic data set.

Finally, we discuss the practical aspects of the application of the proposed method of error prediction and rejection. We argue that credence-based rejection of decisions labeled as unreliable and subsequent application of a relevant repair strategy allows for a reduction of system classification errors to a desired target level. At the same time, such error reduction comes at a cost of running the repair procedures. We show how the benefits and costs of applying a repair strategy can be gauged using the proposed credence evaluation tools.

This chapter is structured as follows. In Section 4.2.1 we discuss the theoretical aspects of single error prediction and probability of a single classifier decision from the subjective Bayesian perspective. In Section 4.2.2 we discuss the extension of the $Q - stack$ framework for credence estimation and error prediction. In Section 4.2.3 we propose a set of evaluation criteria for credence estimates. In Section 4.3 we provide an example of error prediction using proposed method using a synthetic data set. Section 4.4 discusses practical relevance of the proposed error-prediction techniques in autonomous and supervised classification systems. Section 4.5 concludes this chapter with a summary of the presented findings.

4.2 Predicting single classifier errors

4.2.1 Probability of error

Assume two class-conditional distributions of evidence, $p(\mathbf{e}|A)$ and $p(\mathbf{e}|B)$, where A and B are class labels corresponding to arbitrary observation-generating stochastic processes. Examples of such arbitrary distributions for one-dimensional evidence $\mathbf{e}$ are shown in Figure 4.1, but the actual dimensionality of the classification problem in general case is limited only by the number of components of the evidence vector. Consider a classifier that divides the evidence space into two disjoint subspaces by a decision surface Ψ. The decision surface lives in a space whose dimensionality is equal to that of the dimensionality of the feature vector. Classification of observation $\mathbf{e_0}$ is done according to:

$$\Psi(\mathbf{e_0}) > 0 \Rightarrow B(\mathbf{e_0})$$
$$\Psi(\mathbf{e_0}) \leq 0 \Rightarrow A(\mathbf{e_0})$$

Let us consider all observed evidence vectors that satisfy $\Psi(\mathbf{e_0}) \leq 0$. For a given observation whose evidence vector is $\mathbf{e_0}$ we wish to know what is the probability that $\mathbf{e_0}$ is assigned an erroneous class label.

Following the subjective Bayesian interpretation of probability, as discussed in Chapter 2, we define the probability of single error (error credence) $R(\mathbf{e_0})$ using Bayes formula [42] as follows:

$$R(\mathbf{e_0}) = 1 - P(F|\mathbf{e_0}) = 1 - \frac{p(e = \mathbf{e_0}|F) \cdot P(F)}{p(e = \mathbf{e_0})}$$

Here, F denotes an event of misclassification. The alternative event, when the classifier takes a correct decision, is denoted as C^*. Consider the conditional distributions of evidence which has been correctly or incorrectly classified, $p(\mathbf{e}|C)$ and $p(\mathbf{e}|F)$. It is clear that the error-conditional distributions depend on which side of the decision boundary Ψ lies the observation $\mathbf{e_0}$ and that

*The events C and F can be treated as classes, as in [87].

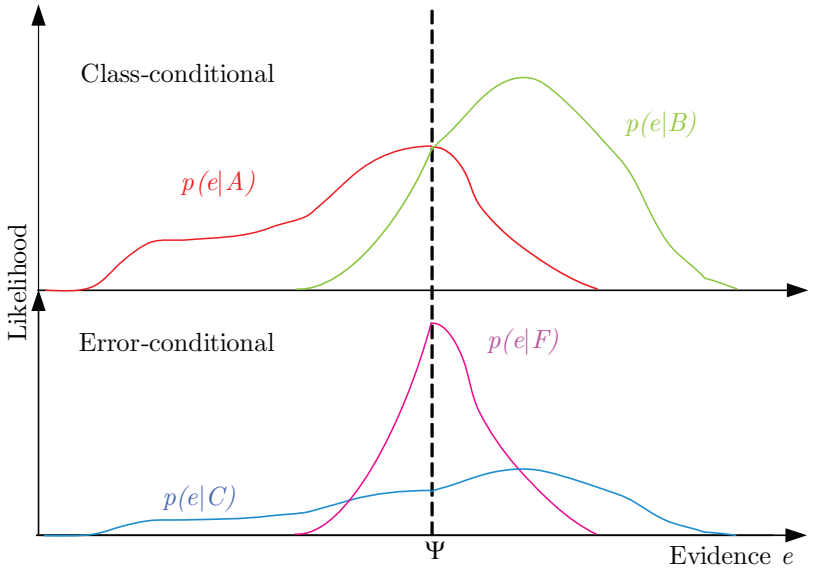


Figure 4.1: Class-conditional (upper) and error-conditional (lower) distributions of evidence **e**.

they are strictly linked with the class-conditional distributions $p(\mathbf{e}|A)$ and $p(\mathbf{e}|B)$. Example of error-conditional distributions are shown in Figure 4.1. For $\Psi(\mathbf{e_0}) \leq 0$ we have

$$\Psi(\mathbf{e}) \leq 0 \Rightarrow p(\mathbf{e_0}|F) = \frac{p(\mathbf{e_0}|B)}{\int_{-\infty}^{\Psi} p(\mathbf{e}|B)de}$$

Obviously in the assumed range of **e**

$$\int_{-\infty}^{\Psi} p(e|B)de = 2P(F)$$

and therefore, again for $\Psi(e) \leq 0$ and considering that the prior class probabilities are assumed equal (Section 2.3.1), $P(A) = P(B) = \frac{1}{2}$

$$P(F|\mathbf{e_0}) = \frac{p(\mathbf{e_0}|B) \cdot P(F)}{2P(F) \cdot p(\mathbf{e_0})} = \frac{\frac{1}{2}p(\mathbf{e_0}|B)}{p(\mathbf{e_0})} = \frac{p(\mathbf{e_0}|B)}{p(\mathbf{e_0}|A) + p(\mathbf{e_0}|B)} = P(B|\mathbf{e_0}). \tag{4.1}$$

Note that mathematically speaking the possible bias of the base classifier that produces scores x is immaterial - the bias becomes corrected by assigning a new decision boundary Ψ in the evidence space. This effect is a characteristic feature of the stacking architecture [122, 165].

Analogically, for $\Psi(\mathbf{e}) > 0$, $P(F|\mathbf{e_0}) = P(A|\mathbf{e_0})$. Acknowledging that $P(A|\mathbf{e_0}) + P(B|\mathbf{e_0}) = 1$ and $P(C|e = e_0) + P(F|e = e_0) = 1$, we compute the credence in a single decision correctness

$$\begin{aligned} \forall_{\Psi(\mathbf{e_0})\leq 0}\ R(\mathbf{e_0}) &= 1 - P(F|e = \mathbf{e_0}) = P(A|e = \mathbf{e_0}) \\ \forall_{\Psi(\mathbf{e_0})>0}\ R(\mathbf{e_0}) &= 1 - P(F|e = \mathbf{e_0}) = P(B|e = \mathbf{e_0}) \end{aligned} \tag{4.2}$$

If all available evidence pertinent to an individual decision has been aggregated in the evidence vector $\mathbf{e_0}$ then the classification decision is as close to optimality as the chosen classifier and its parameters allow. In this situation one cannot do better at classifying the given observation unless

new evidence becomes available, and the credence estimate conveys the information about the degree of belief in the correctness of the classification decision. Application dependent, one may be reluctant to take decisions which have low credence $R(e_0)$. If restraining oneself from taking an uncertain classification decision is a viable option in the given application scenario, a straightforward way of filtering out the uncertain decisions is to discard all decisions whose credence estimate falls below a preset value of ϑ, which we shall refer to as *credence threshold*. The procedure is analogous to the well-known *error-reject tradeoff* first introduced by Chow in [32], and then followed and expanded on in [34, 35, 52, 54, 55].

The procedure of discarding uncertain decisions is represented graphically in Figure 4.2. In the upper graph two class-conditional distributions are exemplified by Gaussians $p(e|A)$ and $p(e|B)$. The optimal classifier is quadratic with two decision thresholds $\tau_1 < \tau_2$:

$$\begin{aligned} &\Psi(\tau_1 < \mathbf{e} < \tau_2) > 0 \\ &\Psi(\{\mathbf{e} < \tau_1\} \cup \{\tau_2 < \mathbf{e}\}) < 0 \end{aligned} \tag{4.3}$$

The corresponding posteriors $P(A|\mathbf{e})$ and $P(B|\mathbf{e})$ are shown in the lower graph. According to Equation 4.2, for an arbitrarily chosen credence threshold ϑ the decisions, whose credence $R(\mathbf{e}) >= \vartheta$ all lie in the area marked by blue color. Corresponding locus of discarded decisions whose $R(e) < \vartheta$ are marked by red color in the upper graph. Note that the zones of $\mathbf{e}$ in which decisions are being discarded due to their low credence surround the decision thresholds τ_1 and τ_2, which agrees with the reasoning given in Section 3.2.

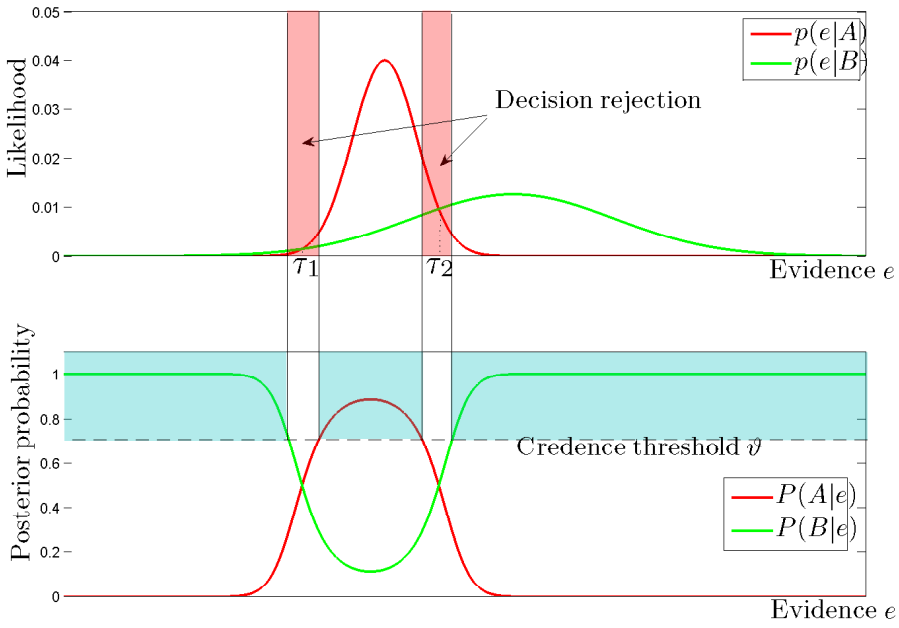


Figure 4.2: Graphical representation of the credence-based error rejection mechanism.

It must be noted here that while the Gaussian models similar to those presented in Figure 4.2 illustrate well how the probability of single classification decision is computed, such models may not be the best choice in practical applications. Note that unless the variances of $p(\mathbf{e}|A)$ and $p(\mathbf{e}|B)$ are

equal, $\sigma_A^2 = \sigma_B^2$, the equation $p(\mathbf{e}|A) = p(\mathbf{e}|B)$ always has two unequal roots at τ_1 and τ_2. In this situation it follows that $\sigma_A^2 \ll \sigma_B^2 \Rightarrow \forall_{\mathbf{e}} P(A|\mathbf{e}) \ll 1$, . Therefore the maximal attainable credence in decisions in favor of class A is always much less than that of the alternative class, $R(\mathbf{e}) \ll 1$ while $\forall_{\{\mathbf{e} \gg \tau_2\} \cup \{\mathbf{e} \ll \tau_1\}} P(B|e) \rightarrow 1$, and one can only obtain certain decisions of $R \approx 1$ for one of the classes (here: class B), and not the other one. This is the case in the example shown in Figure 4.2.

If the class-conditional evidence distributions of $p(e|A)$ and $p(e|B)$ indeed are Gaussian with large variance differences such a result is correct from the viewpoint of Bayesian decision theory [42]. However, in many cases the observed distributions are only approximately Gaussian and the fact that $p(\mathbf{e}|B) > p(\mathbf{e}|A)$ for $\Psi(\mathbf{e}) < \tau_1$ is a mere modeling artifact not supported by the observation data. As a good example may serve here the match versus non-match score histograms and distributions for fingerprint verification from figure 1 of [149]: the estimated curves of class-conditional score distributions intersect in two distinct locations in spite of the fact that the data histograms do not appear to support such pdf estimation. Here, an attempt to estimate the decision credence for the match scores would result in the posterior probability necessarily much smaller than unity. Unless the observed data or prior domain knowledge suggest otherwise, such situation should be avoided by choosing an appropriate data model, for instance by using data histograms or non-parametric distribution estimators [42, 152].

This leads us to another theoretical speculation with important consequences for credence estimation. A pair of Gaussian class-conditional evidence models, as exemplified by the curves in Figure 4.2, are guaranteed to allow for an estimation of decision credence that will obey $\forall_{\mathbf{e}}\ 0 < R(\mathbf{e}) < 1$. However, arbitrary models of $p(\mathbf{e}|A)$ and $p(\mathbf{e}|B)$ may take a value of 0 for certain range of evidence values $\mathbf{e}$. Naturally according to Equations 4.1 and 4.2

$$\begin{aligned} &\forall_{\Psi(\mathbf{e_0}) \leq 0}\ p(e = \mathbf{e_0}|B) = 0 \Rightarrow R(\mathbf{e_0}) = 1 \\ &\forall_{\Psi(\mathbf{e_0}) > 0}\ p(e = \mathbf{e_0}|A) = 0 \Rightarrow R(\mathbf{e_0}) = 1, \end{aligned} \tag{4.4}$$

which is a desired and intuitively justified result. Now, a problematic situation arises when $p(\mathbf{e_0}|A) = p(\mathbf{e_0}|B) = 0$. For such observation $R(\mathbf{e_0})$ is a mathematical singularity, with undefined value according to Equation 4.1. Typically, $p(\mathbf{e_0}|A) = p(\mathbf{e_0}|B) = 0$ means that $\mathbf{e_0}$ is simply not accounted for by the models. Dependent on the application considered, one may choose to either arbitrarily assign:

- $R(\mathbf{e_0}) = 0.5$. This formulation assigns the credence equal to the level of chance to $R(\mathbf{e_0})$, which means that there is no reason to favor one decision over its alternative since there is no evidence for such choice. In classification problems where prior probabilities are unequal, the decision could be taken based on the priors only. Such formulation agrees with the viewpoint that degree of support and degree of belief are to be considered as separate concept [72]. Here, $R(\mathbf{e_0}) = 0.5$ expresses the degree of belief, since the degree of support is zero. Similar reasoning about evidential support can be found to be at the roots of the Dempster-Schafer theory [135].

- $R(\mathbf{e_0}) = 1$. Consider the curves shown in Figure 4.1. Let us assume that $\forall_{\mathbf{e_0}: \Psi < \kappa < \mathbf{e_0}} p(e = \mathbf{e_0}|A) = p(\mathbf{e_0}|B) = 0$. Since for $\mathbf{e_0} > \kappa$ also $e > \Psi$ it might be pragmatically argued that if any class label is to be assigned then it should be B. Such a classification decision, in fact not supported by Bayesian reasoning, is an example of non-monotonic *default reasoning* [135]. Since from the decision standpoint it is most important on which side of the decision hyperplane falls $\mathbf{e_0}$ then the fact itself constitutes evidence to support the decision. In the absence of any evidence at all to support alternative decision, $R(\mathbf{e_0}) = 1$.

Note that choosing $R(\mathbf{e_0}) = 0$ is not a valid option. Regardless of the actual class label assigned to e_0 (as discussed above), choosing $R(e_0) = 0$ would mean that actually the opposite class should be chosen. Moreover, it would mean that the opposite decision is absolutely certain - an obvious fallacy in the light of actual absence of any evidence to support such, or in fact any decision.

4.2.2 $Q - stack$ for credence estimation

Note that in Equation 4.2 the terms $P(\mathbf{e_0}|A)$ and $P(\mathbf{e_0}|B)$ describe class-conditional posterior probabilities for classes A and B, respectively. Since the assigned class label depends on the sign of $\Psi(\mathbf{e_0})$ the term $P(C|\mathbf{e_0})$ is always the posterior probability computed for $\mathbf{e_0}$ in respect to the selected class label. Posterior probabilities are a natural output of a naive Bayes classifier [42, 111].

Consider Figures 3.13, 3.14 and 3.15. As mentioned in previous Sections, the stacked classifier in the $Q - stack$ architecture may be any classifier that is appropriate for the structure of the evidence data. If the chosen stacked classifier is a Bayes classifier then on the output of $Q - stack$ it will generate class-conditional posterior probabilities. These probabilities computed for the selected class are the credence estimates in the correctness of single classification decisions, in accord with Equation 4.2.

The similarity of the proposed approach to the classical Chow's error-reject tradeoff, also based on estimates of posterior probabilities, calls for a few words of comment. The classical Chow's approach is relying on an exact estimation of posterior probabilities by the baseline classifier [55]. Due to the classifier bias, these estimates may also be inaccurate [155]. In the proposed scheme of $Q - stack$ the posterior probabilities are estimated not by the baseline, but by the second-level, stacked classifier. The benefits of this, in comparison to the classical Chow's error-reject tradeoff, are two-fold:

- A properly trained stacked classifier ensemble has the property of reducing the classifier bias, which translates directly into reduced error in estimation of posterior probabilities.

- The stacked classifier takes into account the information about the quality of classified signals, which is invisible to the baseline classifier. For this reason, given relevant quality measures, the stacked classifier will produce more accurate posterior probability estimates than the baseline classifier.

4.2.3 Evaluation criteria for credence estimates

Evaluating credence estimates is not an obvious task. Since credences as seen from the subjective Bayesian perspectives are degrees of belief in the occurrence of a given probabilistic event, they express a priori probability of an event that has not happened yet, and may not actually have any precedents. For this very reason single event probability is not accounted for by classical or frequentist interpretation of probability. Therefore the actual evaluation of a single credence estimate is not actually feasible. However, credence estimates can be evaluated a posteriori, for a given testing data set. The adequacy of credence estimation can be evaluated using following criteria:

1. **Monotonicity**: The total accuracy of the classifier decisions labeled as dependable for a given credence threshold ϑ must follow the tendency of the credence threshold: the higher the threshold, the lower the classification error ER of the remaining decisions after the credence thresholding:

$$\forall_{\vartheta_1,\vartheta_2:\vartheta_1<\vartheta_2} ER(\vartheta_1) \geqslant ER(\vartheta_2) \tag{4.5}$$

2. **Accountability**: Since the credence of a decision is in fact an estimate of probability of a single random event, it is not evident how to gauge its accuracy. However, it can be reasonably expected that if the estimates are accurate over a large enough set of $m = 1, 2, ..., M$ sample decisions obtained for evidence vectors $\mathbf{e}_m$, their average credence will account for the actual observed accuracy over the sample set. The average credence $\overline{R}$ is computed as follows:

$$\overline{R} = \frac{1}{M} \sum_{m=1}^{M} R(\mathbf{e}_m), \tag{4.6}$$

where M is the number of available testing samples. Formally, the difference between the average credence $\overline{R}$ over the testing data set and the actual accuracy over this data set must be minimized:

$$\delta = \overline{R} - AC \approx 0 \tag{4.7}$$

In Equation 4.7, the classifier accuracy is given by

$$AC = \frac{1}{M} \sum_{m=1}^{M} \neg \left((D_m \wedge GT_m) \vee (\neg D_m \wedge \neg GT_m) \right), \tag{4.8}$$

where D_m is a binary decision obtained for $\mathbf{e}_m$, and GT_m is the true binary class alignment of $\mathbf{e}_m$.

The proposed evaluation criteria for credence estimation agree with an intuitive understanding of the concept of credence. Namely, informally stated they demand that decisions deemed more credible should turn out to be indeed more accurate.

For data sets where one of the classes is represented by significantly more samples than another estimating the classification performance in terms of total accuracy can be misleading - error for the over-represented class will dominate the total error estimate. unbalanced data sets are frequent in biometric identity verification, where the class of imposter scores is typically over-represented.In this situation frequently $HTER$ rather than AC is used for performance evaluation [120]. The same problems with unbalanced sets concern the accountability criterion for credence estimates. In order to compare the mean credences with HTER over an unbalanced data set a modification of Equation 4.7 applies:

$$\delta = \overline{R_w} - (1 - HTER) \approx 0, \tag{4.9}$$

where $\overline{R_w} = \frac{1}{2}(R_A + R_B)$. Here, R_A and R_B are mean credence estimators for classes A and B, respectively.

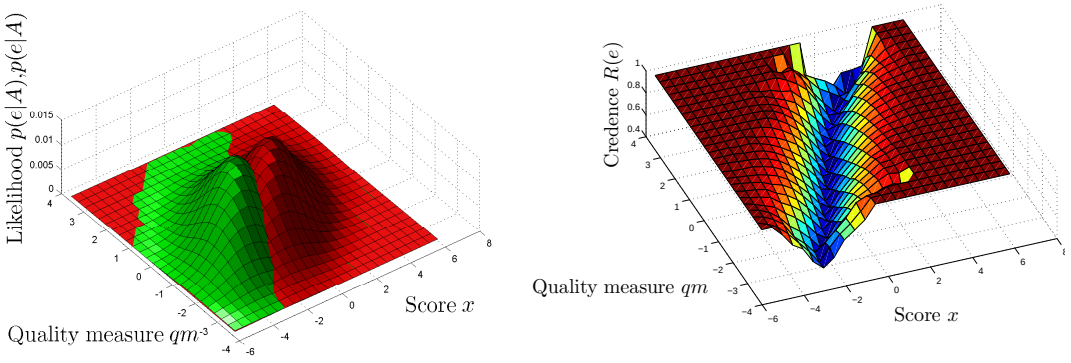
4.3 Examples: synthetic data

4.3.1 Gaussian case - equal distributions

In order to demonstrate the principles of error prediction using credence estimates we use the same example as given in Section 3.8. The parameters of the data- and noise generating processes are given by Equations 3.20 and 3.22. Recall that one of the classifiers used in this experiment was a Bayesian classifier - its decision boundaries are shown in Figures 3.19(b) and 3.20(b). In the experiment shown in this section we will use the class-conditional posterior probabilities $p(A|\mathbf{e})$ and $p(B|\mathbf{e})$ as estimated by the Bayes classifier in order to find the credence estimates for each observation, as given by Equation 4.2.

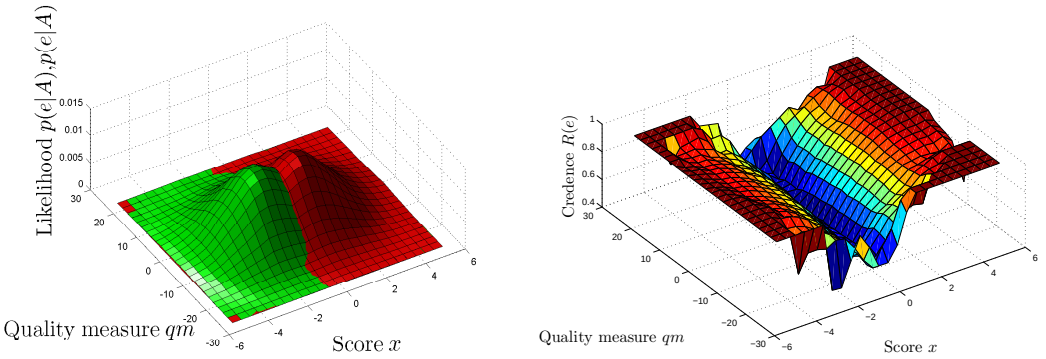
In the example given in Section 3.8 we have analyzed the impact of the uncertainty in measuring the quality measure qm on the classification accuracy in the $Q-stack$ architecture. This uncertainty was modeled by adding white Gaussian noise of preset variance σ_d^2 to the measured quality measures. In the experiment reported in this section we demonstrate the impact of this uncertainty on the error prediction properties of credence estimates. For the purpose of the demonstration we have selected two extreme cases: $\sigma_d^2 = 0$ which represents the situation where quality measures represent the noise contamination of the data with no uncertainty, and $\sigma_d^2 = 50$ represents a very noisy measurement of the signal quality. The corresponding correlation coefficients ρ between the scores and quality measures can be found in Table 3.8.

Figure 4.3 gives a graphical representation of the example classification problem, for $\sigma_d^2 = 0$ and Figure 4.4 shows the same for $\sigma_d^2 = 50$. Figures 4.3(a) and 4.4(a) show 2D histograms of testing evidence $\mathbf{e}|A$ and $\mathbf{e}|B$ for 1000 data points per class. Figures 4.3(b) and 4.4(b) show the credence estimates for class decisions computed according to Equation 4.2.



(a) Distributions (b) Credence

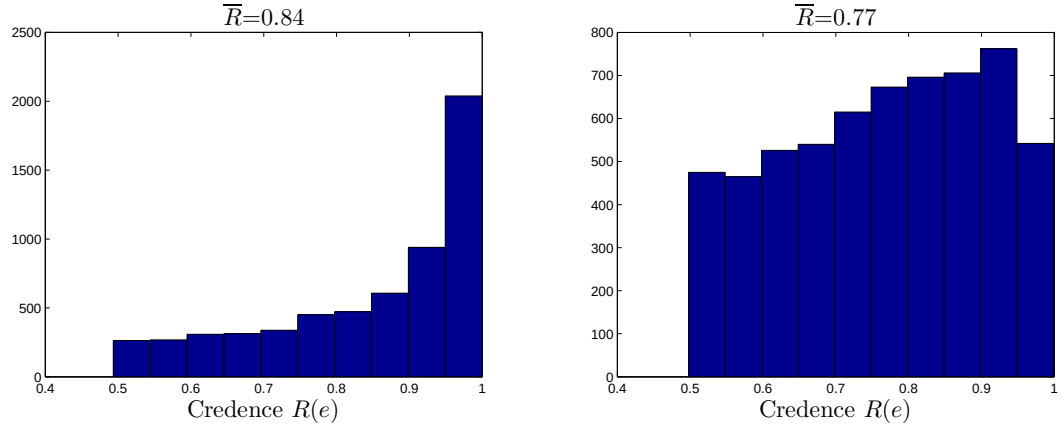
Figure 4.3: Class-conditional evidence distributions $p(\mathbf{e}|A)$ and $p(\mathbf{e}|B)$ and computed credence estimates $R(e)$. Quality measures taken at $\sigma_d^2 = 0$.



(a) Distributions (b) Credence

Figure 4.4: Class-conditional evidence distributions $p(\mathbf{e}|A)$ and $p(\mathbf{e}|B)$ and computed credence estimates $R(e)$. Quality measures taken at $\sigma_d^2 = 50$.

It is easy to notice that when $\sigma_d^2 = 0$ the escarpments of the surface plot of credence $R(\mathbf{e})$ are much steeper than it is the case for $\sigma_d^2 = 50$. In order to better visualize the distribution of the credence values in each case we show the corresponding histograms in Figure 4.5. As the reader will readily notice, the histogram for the uncertain quality measure estimates is visibly more flat. In the case of $\sigma_d^2 = 0$ majority of the decisions are taken at a high level of confidence which is reflected in credence values clustered in the proximity of $R(\mathbf{e}) = 1$. This is not the case for $\sigma_d^2 = 50$: since the quality measurement is not dependable, large number of classification decisions are taken at low level of confidence, represented by low credence estimates.



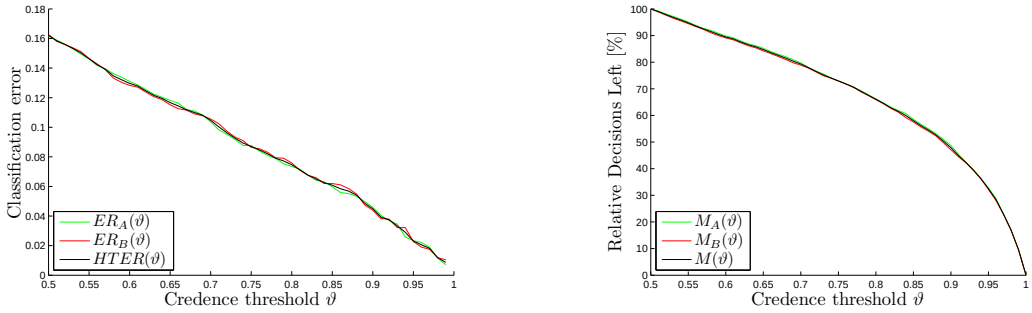
(a) $\sigma_d^2 = 0$ (b) $\sigma_d^2 = 50$

Figure 4.5: Histograms of credence estimates $R(\mathbf{e})$ for quality measures taken at $\sigma_d^2 = 0$ and $\sigma_d^2 = 50$.

Figures 4.6(a) and 4.7(a) show the accuracy of classification decisions left after the procedure of credence thresholding, as a function of the credence threshold, $ER(\vartheta)$. The curves obtained for $\sigma_d^2 = 0$ shown in Figure 4.6(a) and for $\sigma_d^2 = 50$ in Figure 4.7(a) allow for the evaluation of the credence estimates according to the monotonicity criterion. Evidently, as the credence threshold increases from $\vartheta = 0$ towards $\vartheta = 1$, the error rates for the decisions remaining after credence thresholding decrease, thus satisfying the criterion expressed by Equation 4.5

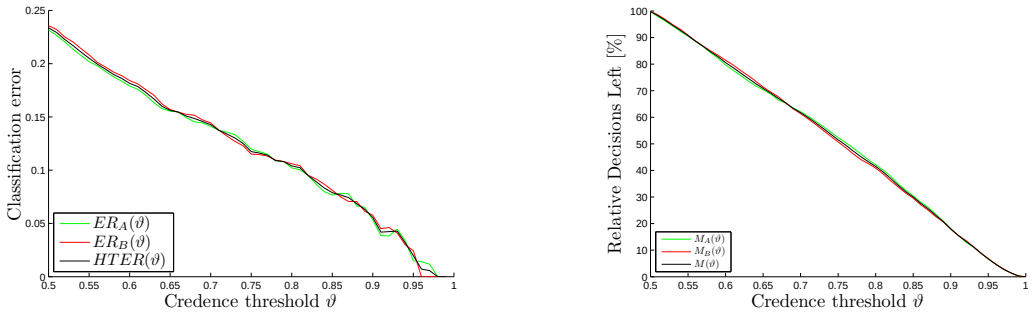
If decisions whose credence estimates falls low and below the credence threshold are to be discarded, then the shape and steepness of the credence histogram predicts how many decisions will be discarded for the values of ϑ equal to the histogram bin centers. if a histogram is flat that suggests that comparable volume of decisions will be discarded at every credence threshold ϑ. This is confirmed by Figures 4.6(b) and 4.7(b). Note that in figure 4.7(b) the curve that represents the percentage of decisions kept after credence thresholding is approximately linear. Due to the quality measure uncertainty introduced by non-zero σ_d^2 more decisions had to be discarded than it is the case of accurate qm measurement.

An evaluation of the presented method of estimating decision credence according to the accountability criterion is shown in Table 4.1 for $\sigma_d^2 = 0$ and in Table 4.2 for $\sigma_d^2 = 50$. The experiments described above were repeated 100 times, and for each experiment run the total accuracy and the mean credence $\overline{R}$ computed according to Equation 4.6 was recorded. The mean and standard deviation of the absolute difference $|AC - \overline{R}|$ is a measure of δ, Equation 4.7. Small value of the asymptotic p-value p_κ computed using the two-sample two-sided Kolmogorov-Smirnov test [37] shows that with very high level of certainty it can be concluded that AC and $\overline{R}$ are samples drawn from the same



(a) (b)

Figure 4.6: Error rejection curves for quality measures taken at $\sigma_d^2 = 0$.



(a) (b)

Figure 4.7: Error rejection curves for quality measures taken at $\sigma_d^2 = 50$.

	$AC = 1 - HTER$	$\overline{R}$	δ	p_κ
mean	$\overline{AC} = 84.13\%$	$\overline{\overline{R}} = 84.18\%$	$\overline{\lvert AC - \overline{R}\rvert} = 0.74\%$	$5.81 \cdot 10^{-4}$
standard deviation	$\sigma_{AC} = 0.94\%$	$\sigma_{\overline{R}} = 0.33\%$	$\sigma_{\lvert AC-\overline{R}\rvert} 0.56\%$	

Table 4.1: Evaluation of the credence estimates according to the accountability criterion, for $\sigma_d^2 = 0$.

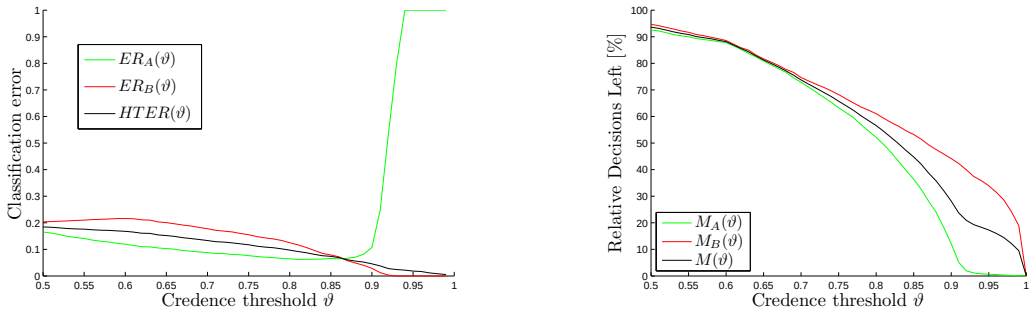
	$AC = 1 - HTER$	$\overline{R}$	δ	p_κ
mean	$\overline{AC} = 76.1\%$	$\overline{\overline{R}} = 76.16\%$	$\overline{\lvert AC - \overline{R}\rvert} = 0.81\%$	$6.91 \cdot 10^{-4}$
standard deviation	$\sigma_{AC} = 0.9\%$	$\sigma_{\overline{R}} = 0.5\%$	$\sigma_{\lvert AC-\overline{R}\rvert} 0.67\%$	

Table 4.2: Evaluation of the credence estimates according to the accountability criterion, for $\sigma_d^2 = 50$.

population and hence the accountability criterion expressed by Equation 4.7 is met.

4.3.2 Gaussian case - unequal marginal score distributions

In the previous example we have been considering an idealized case where $\sigma_A^2 = \sigma_B^2$. In such case both class errors $ER_A(\vartheta)$ and $ER_B(\vartheta)$ monotonically decline as the credence threshold approaches $\vartheta \approx 1$. However, we have also mentioned in Section 4.2.1 that for $\sigma_A^2 \neq \sigma_B^2$ that will not be the case. Here we return to this topic in order to illustrate with a synthetic data example the performance of credence estimates as an error predictor when the class variances of normal class-conditional evidence distributions are unequal. For this, we use the same distribution parameters given by Equation 3.20 except $\sigma_A^2 = 2 \neq \sigma_B^2$. Following the procedures described in Section 4.2.1 we draw the error prediction curves for $ER_A(\vartheta)$, $ER_B(\vartheta)$ and $HTER(\vartheta)$, and the curves giving the relative number of discarded classification decisions $M_A(\vartheta)$, $M_B(\vartheta)$ and $M(\vartheta)$. These graphs are shown in Figure 4.8.



(a) (b)

Figure 4.8: Error rejection curves for quality measures taken at $\sigma_d^2 = 0$ for unequal variances $\sigma_A \neq \sigma_B$.

The immediately apparent difference between the graphs in Figure 4.6 and in Figure 4.8 is the shape of $ER_A(\vartheta)$. Instead of maintaining a monotonically decreasing shape, in Figure 4.8 the $ER_A(\vartheta)$ decreases towards a certain value of ϑ, and as the credence threshold continues to grow

	$AC = 1 - HTER$	$\overline{R}$	δ	p_κ
mean	$\overline{AC} = 79.68\%$	$\overline{\overline{R}} = 79.61\%$	$\overline{\lvert AC - \overline{R}\rvert} = 0.74\%$	0.047
standard deviation	$\sigma_{AC} = 0.86\%$	$\sigma_{\overline{R}} = 0.41\%$	$\sigma_{\lvert AC-\overline{R}\rvert}0.57\%$	

Table 4.3: Evaluation of the credence estimates according to the accountability criterion, for $\sigma_A^2 = 2 \neq \sigma_B^2$ and $\sigma_d^2 = 0$.

towards 1 the $ER_A(\vartheta)$ grows as well. This effect is a consequence of the fact that maximal credence estimate that can be reached for class A is smaller than what can be obtained for class B.

The evaluation of the credence estimates according to the accountability criterion is given in Table 4.3, again for 100 repetitions of the experiment. Despite the fact that the behavior of $ER_A(\vartheta)$ is erratic above certain value of ϑ, the value of δ remains small and the accountability criterion is still met, albeit at a higher value of the asymptotic p-value p_κ than it was the case when $\sigma_A^2 = \sigma_B^2$ (Tables 4.1 and 4.2). This observation shows that indeed the erratic behavior of $ER_A(\vartheta)$ is relevant to only a few decisions.

Let us denote the maximal credence value that can be reached for class A as R_{max}. As the credence threshold ϑ approaches R_{max}, most potentially erroneous decisions for class A have been discarded. This effect is confirmed by the shape of $M(\vartheta)$ in figure 4.8 (b) which reaches values close to 0 for $\vartheta \approx R_{max}$. As the credence thresholding continues for $\vartheta > R_{max}$ the remaining decisions assigned to class A are actually misclassifications, and hence the growing shape of $ER_A(\vartheta)$.

In the case of bivariate Gaussian distributions considered in the examples discussed here it is easy to analytically find the value of R_{max} beyond which the credence thresholding for the particular class ceases to return lower classification error rates. R_{max} can be computed as given by Equation 4.10.

$$R_{max} = R(\frac{\sigma_A^2\mu_B - \sigma_B^2\mu_A}{\sigma_A^2 - \sigma_B^2}) \tag{4.10}$$

However, in the case of multi-variate normal distributions, their mixtures, and non-Gaussian class conditional evidence distributions analytical computation of R_{max} easily becomes impractical and often not possible (for instance when the distributions are estimated in a non-parametric fashion). In this case, the maximal admissible value of credence for each class R_{max} can be found from graphs similar to those shown in Figure 4.8, drawn for the available training data. The sought value of R_{max} is then indicated by the value of ϑ for which the class error curve reaches its minimum.

4.4 Practical relevance

The presented method of classifier error prediction and rectification was developed with the intention of reducing classification errors by discarding potential misclassifications from the total decision pool. A decision, whose credence estimate falls below a present threshold ϑ is considered unreliable and a repair strategy is applied. The essence of a repair strategy is that upon encountering an unreliable decision the classification system requests a new presentation of data for subsequent classification. Depending on the application, a repeated presentation of data from the same or from a different modality can be requested, and the classification procedure starts from the beginning to conclude with a new decision and its credence estimate. If the new credence estimate is above ϑ the repair strategy is successful. Again, depending on the application and classifier architecture used, the system may continue to re-request and re-classify new data presentations until the repair loop arrives at a dependable decision. An example of a similar sequential repair strategy applied to

biometric speaker verification is discussed in [129], where a decision deemed unreliable is followed by a repeated presentation of speech data. As a result of the application of this strategy a significant gain of classification was reported. We do not pursue the topic of developing more sophisticated repair strategies any further in this thesis.

Of course each iteration of any repair strategy comes at the cost of extended classification time, allocation of additional computing power and user inconvenience and the benefit of using a repair strategy must outweigh this extra costs. It is therefore of critical importance to estimate potential gains and losses which are expected to result from rejecting unreliable classification decisions. Obviously the gain is the expected improvement in classification accuracy. Since repairing each unreliable decision is expensive, the expected cost is the projected volume of decisions that need to be rejected and consequently repaired. Here we show an example how these can be gauged in the proposed scheme of credence-based error prediction.

Consider sample curves shown in Figure 4.9. The already familiar graph shown in Figure 4.9(a) shows the classification error of a system as a function of the credence threshold ϑ, obtained using a development data set of known groundtruth. The generic term *error*, ER, in this graph denotes either class-specific error, total error, $HTER$, or any other applicable error measure. Figure 4.9(b) shows the number of decisions remaining after credence thresholding, as described in the examples from Section 4.3.1.

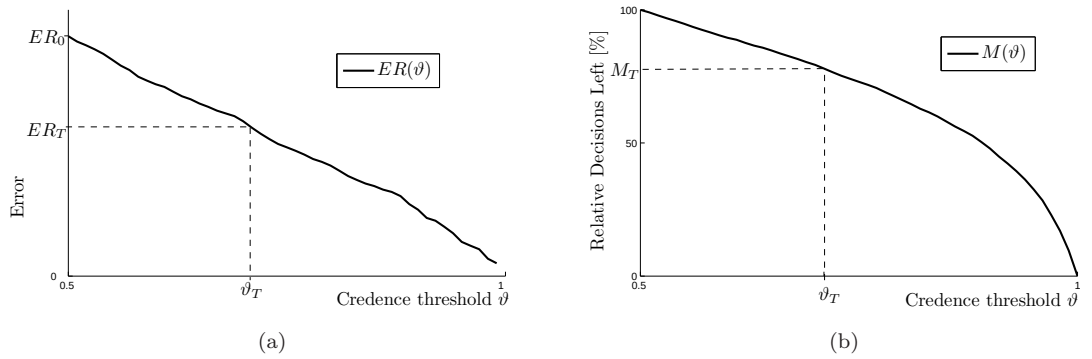


Figure 4.9: Improving classification performance using credence-based error prediction and rejection. (a) Finding credence threshold ϑ_T corresponding to the desired target error rate ER_T, and (b) finding the number of decisions $100\% - M_T$ to be discarded in order to arrive at the target error rate ER_T.

Let ER_0 be the original error of the entire decision pool containing 100% of classification decisions. Assume that the obtained error of ER_0 is excessive for the considered application, and it is necessary to reduce the target error rate to $ER_T < ER_0$. From Figure 4.9(a) it is possible to find a value of credence threshold ϑ_T, which guarantees that the classification error for decisions remaining after the credence thresholding procedure will be ER_T, as required. At the same time, figure 4.9(a) reveals the cost associated with reaching the target classification error. The value of ϑ_T defines M_T, the percentage of decisions remaining in the decision pool after the unreliable decisions have been discarded in order to achieve ER_T. Thus the cost of achieving the target error reduction can be gauged by the cost of applying the repair strategy to $100\% - M_T$ decisions.

Proposed method of credence estimation is also relevant to supervised biometric authentication. Most commonly existing, large-scale applications of biometrics are not fully autonomous but require human supervision. Beside actually improving the classification accuracy, the methods proposed in

this paper can be applied in such systems as a decision support technology. We refer to this scenario as to the Identity Verification Assistant (IVA). An argument in favor of the use of the biometric verification systems in the IVA scenario is that identity verification, depending on the importance of the application, is bound to have legal implications. As of today, it is hard to envision an automatic biometric verification system taking fully autonomous decisions in, for instance, immigration control. There is always a need of a physical person who is responsible for the decisions taken regarding identity verification, and therefore large-scale biometric systems are likely to be, at least partially, supervised. A biometric verification system that would output only a binary yes/no decision will be of little use in the IVA scenario, unless it would be certain to commit no mistakes (which is not the case). Since errors are inevitable, it is necessary that next to the actual binary decision, the system produces an estimate of decision credence. The dependability estimates must be expressed in the terms that are intuitively understood by the supervising persons and the probabilistic formulation of credence meets this requirement [116].

4.5 Summary

In this chapter we have presented a discussion of the problem of estimating credence, or degree of belief in the correctness of single classification decisions. Since single event probability is a concept undefined in many probabilistic systems, including the classical interpretation of probability, we adapted the subjective Bayesian interpretation in order to derive a predictor of misclassifications. Consequently we have derived the formulation for computing decision credence based on available evidence. The proposed method of estimating single decision credence is a functional extension of the $Q - stack$ scheme of classification with class-independent quality information, presented in Chapter 3. Consequently, the proposed method of error prediction also makes use of available quality measures. The credence estimates are found by applying a Bayesian stacked classifier in the $Q - stack$ framework. The posterior probabilities produced by the Bayes classifier during evidence classification are the sought credence estimates. These estimates are further applied to detect and discard unreliable decisions from the total decision pool. We have discussed the relationship between the proposed method and the classical Chow's error-rejection tradeoff. We have shown that the benefits in using the $Q - stack$ approach come from the inclusion of quality measures in the process of credence estimation, and in the use of the stacking classifier ensemble. A stacked classifier is known to reduce the baseline classifier bias, and consequently the error in estimation of the decision credence. More accurate estimation of credence in individual classification decisions translates into improved error prediction performance of the $Q - stack$ ensemble in comparison with the baseline classifier.

We have proposed evaluation criteria for the presented credence estimates, which are formalizations of an intuitive requirement that more credible decisions must turn out to be more accurate. We have demonstrated the performance of the proposed error prediction method using a synthetic data set. We have also shown results which stress the role of relevant quality measures in accurate error prediction. We have discussed how credence estimates can be used to identify potentially erratic decisions which can be rectified using a repair strategy. We have proposed a procedure which allows for gauging the benefits and costs of applying a repair strategy in a practical pattern classification problem.

Evaluation - biometric identity verification

5

5.1 Introduction

This chapter is dedicated to demonstrating how quality measures help better classify real multimodal biometric data. For this purpose, we apply the theoretical framework of $Q - stack$ proposed in Chapter 3 to the task of biometric identity verification using face and fingerprint data. Experiments reported here are conducted using data from a real multi-modal database - Biosec [47]. The Biosec database contains data collected for multiple biometric modalities, including face, fingerprint, iris and speech. For reasons discussed in Chapter 1, in this chapter we focus our attention on the face and fingerprint modalities, and on their use in single-, multi-classifier and multimodal scenarios applied to identity verification.

The main goal of the experiments presented here is to evaluate the proposed method of $Q-stack$ as a generalized framework of classification with quality measures, applied to single-, multi-classifier, and multi-modal biometric identity verification. We provide evidence to support following claims:

1. Score-dependent quality measures provide additional dimensions in which a stacked classifier can separate the classes better than the baseline classifier that uses only the similarity scores.

2. Proposed method of $Q - stack$ affords improved biometric verification than corresponding baseline systems in single-, multi-classifier, and multi-modal scenarios.

3. In a multi-classifier system, quality measures need to be dependent on at least one classifier in order to observe the benefits of $Q - stack$.

4. Application of a Bayesian classifier as the stacked classifier allows for accurate prediction of correctness of single classification decisions.

In any experimental evaluation it is of prime importance to clearly define the reference that the presented results are compared against. Since the use of class-independent quality measures in

classification is the main theme of this dissertation, in the experiments reported in this chapter we use the results obtained by the baseline classifiers that do not use quality measures as such reference. Consequently, single baseline classifiers are used as reference to evaluate $Q-stack$ applied to single-classifier systems, multiple-classifier baseline reference applies to $Q-stack$ used in a multiple-classifier architecture, and multi-modal $Q-stack$ is evaluated against a multi-modal baseline system. All baseline systems are classifiers or classifier ensembles that do not use quality measures.

At this point a due explanation is required regarding the choice of the reference systems for a comparative evaluation of $Q-stack$. Since in the subject literature there is prior art that describes methods of using quality measures to improve classification it might seem appropriate to use them as reference. It is not the case. As we have explained in Chapter 3, proposed framework of $Q-stack$ is a generalized approach that encompasses previously proposed methods, that were specifically tuned to a particular application or data set. We have dedicated Section 3.7.2 to a detailed explanation of the fact that examples of heuristic decision functions that use quality measures are in fact more or less accurate approximations of the optimal decision hyperplane drawn in the evidence space. As opposed to the restrictive formulations found in the prior art, proposed method of $Q-stack$ has a theoretical potential of finding an optimal decision boundary in the evidence space for any data structure, given that there exists enough training examples and that the stacked classifier and its parameters are selected optimally. Given that, on a specific data set any of the algorithms previously proposed in the literature may approach optimality, and hence $Q-stack$ could not do any better - both approaches would be approximations of the very same optimal decision hyperplane. At the same time, however, tailor-made heuristic decision boundaries in the evidence space would be likely to perform poorly on data that they were not designed for. In both cases a comparison between $Q-stack$ and these methods would not be fair since they are in fact nothing but a special, heuristic case of $Q-stack$. Also, for a fair comparison of $Q-stack$ with the alternative methods we would have to use exactly the same baseline data and quality measures as the authors of these methods did. This is hardly feasible.

The major difference between the proposed framework of $Q-stack$ and the alternative heuristic approaches is in the way that the used quality measures are treated, and how are they employed in the classification process. Specifically, as opposed to alternative prior methods, in the $Q-stack$ framework the quality measures are nothing but classification features. We wish to reiterate that the proposed method of $Q-stack$ is a generic framework that employs class-independent quality information in order to improve the class separation, and consequently enhance the classification performance. In $Q-stack$, quality measures receive equal treatment as class-selective features. In the experiments described here we demonstrate that quality measures used in such way indeed afford an improved classification of real biometric data.

In Chapter 4 we have presented an extension of the $Q-stack$ framework that allows for accurate error prediction and rejection using credence estimates. In this chapter we apply the proposed credence-based error prediction scheme to the practical classification results obtained using the face and fingerprint data from the Biosec database. An evaluation of the credence estimates is performed according to the criteria set in Chapter 4.

This chapter is structured as follows. Section 5.1.1 describes the used database, experimental setup, and reporting procedures used in this chapter. We provide details concerning the Biosec database and the associated experimental protocol used in the experiments described further in this chapter, with particular attention paid to the face and fingerprint modality. The classifier types used as the second-level stacked classifier in the $Q-stack$ scheme are described and the structure of result reporting is presented.

Section 5.2 focuses on the experiments involving face modality. The baseline classifier systems

are discussed, followed by the description of the proposed face image quality measures used in the experiments, given in Section 5.2.2. In the same section we elaborate on the link between the baseline classifiers and the proposed quality measures, which is of paramount importance for the functionality of the methods proposed in this thesis. The experimental results for classification using the $Q-stack$ scheme and for error prediction applied to the face modality are presented in Section 5.2.3. In Section 5.3 we present the experimental results for the fingerprint modality. The details of the used fingerprint matcher and used quality measures are given, and the dependence link between them is discussed. The experimental results for the fingerprint modality, including classification in the $Q-stack$ framework and the results of error prediction, are presented in Section 5.3.3. In Section 5.4 we show the results of applying $Q-stack$ in a multimodal biometric identity verification architecture involving both face and fingerprint modalities.

Sections 5.5 and 5.6 conclude the chapter with a recapitulation of the outcomes of the experimental work.

5.1.1 The Biosec database and experimental protocol

The part of the Biosec database used in the experiments presented in this chapter consists of face and fingerprint data collected from 200 users. Data collected from 50 of the users' pool is dedicated as *development* (training) set by the associated experimental protocol [5]. Remaining 150 users' data is used for *evaluation* (testing). Since there is no overlap between data or users from the training and evaluation data sets the experimental protocol can be categorized as *open*, meaning that in the intention it should generalize well to unseen data and new users. The full baseline Biosec database consists of face, fingerprint (3 different sensors), iris and voice data collected for each user in four sessions, four samples per user in each session. The examples of data from the database are presented in Figure 5.1. For further details regarding the BioSec database the reader is referred to [5, 47]*.

The choice of the Biosec database over other existing biometric databases is dictated by the fact that it is to the best of our knowledge the only publicly available multimodal, non-chimerical biometric database that contains a relatively large data volume of fingerprint and face images of varying quality. Alternative, publicly available databases provide sufficient volumes of data for only face or fingerprint modalities alone. Building a chimerical database out of data originating from inconsistent individual donors is possible, but the use of chimerical databases has been criticized in the biometric literature [118].

All users in the database have assigned numbers. The entire data set is divided into two disjoint sets, an *evaluation* set consisting of 150 (user numbers 26-175) users, and a *development* set consisting of the remaining 50 (user numbers 1-25 and 176-200) users. No users appear in both development and evaluation sets. As permitted by the experimental protocol [5] we subdivide the development set into a model training subset, denoted as D_1, and a parameter development set, D_2. The subset D_1 consists of data collected from users numbered 1-12 and 176-188 and is used exclusively for training of classifier models, and for training of quality estimator model parameters. The subset D_2 consists of data collected from users numbered 13-25 and 189-200 and is used for tuning the decision thresholds used later in the evaluation. In the experiments reported here, the classification results reported on development data set always refer to the results obtained using the D_2 subset. Classifiers and thresholds trained on the development dataset are then applied to classify the evaluation dataset.

*We wish to thank Prof. Javier Garcia-Ortega and Dr. Julian Fierrez-Aguillar (Universidad Autónoma de Madrid) for making the face part of the Biosec database available for our experiments.

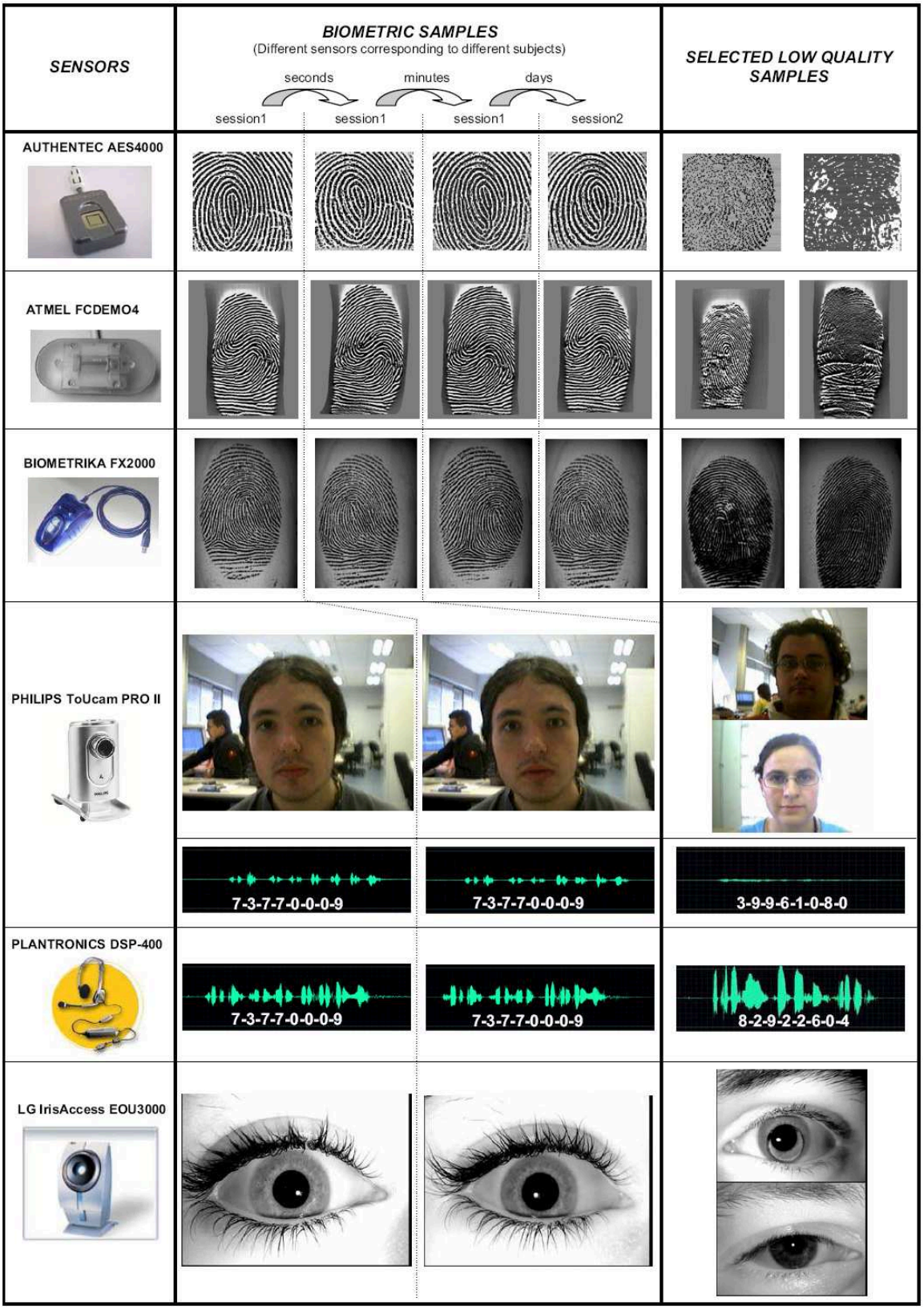


Figure 5.1: Sample records from the Biosec baseline corpus. Image from [47].

number of data points				
	Development			Evaluation
	total	D_1	D_2	
class A: imposter access	1225	703	522	11135
class B: genuine access	800	400	400	2400
class $A + B$: all data	2025	1103	922	13535

Table 5.1: Volume of data per class available in the development and testing sets, per modality, Biosec database.

The experimental protocol defined around the Biosec database declares that all experiments are to be conducted in the one-to-one sample matching scenario. This means that for each individual identity check, one and only one sample is considered a training (gallery) sample and is matched against only one testing sample at a time. The classification task involves distinguishing between the matching scores representing two classes:

- **Class A: imposter claim** - the training sample originates from a different individual than the testing sample, according to the following scheme: for each user the first sample in the first session is matched against the first sample of the same session for every remaining user. Symmetric matches are avoided, i.e. if training sample from user k is matched against a test sample from user l then the same data for user l is not re-used as training sample when matching against the testing sample of user k. The total number of imposter matching scores for class A is:

  Development: 50 subjects × 1 samples × 49 subjects / 2 = 1125 matching scores
  Evaluation: 150 subjects × 1 samples × 149 subjects / 2 = 11175 matching scores.

- **Class B: genuine identity claim** - the training sample originates from the same individual as does the testing sample, according to following scheme [5]: The 4 samples in the first experimental session are matched against the 4 samples in the second session. Resulting total number of genuine matchings generated for each modality:

  Development: 50 subjects × 4 samples × 4 samples = 800 matching scores
  Evaluation: 150 subjects × 4 samples × 4 samples = 2400 matching scores.

Considering the subdivision of the development set into subsets D_1 and D_2, total data volumes for specific data subsets with respect to the classes are given in Table 5.1.

We report the classification performance in terms relative to the data volumes given in Table 5.1. Since the available data set is unbalanced, i.e. classes have uneven sample sizes, in order to be able to properly interpret the percentages it is important to bear in mind what change of classification accuracy results from a change of assigned class label to one single sample. These resulting change in error rates, shown in Table 5.2, give the resolution of the reported classification results.

The multimodal protocol defined for the Biosec database strictly follows the unimodal protocol and respects the actual data collection order: the user, session and sample numbers are the same for each modality.

In the experiments reported in this chapter fingerprint and face modalities are used to demonstrate the use of the proposed framework of classification with quality measures, $Q - stack$, in single-classifier, multi-classifier and multimodal settings. For the face modality two separate face matching algorithms are used that operate on the same sample face images. Combination of the two matchers is used as an example of $Q - stack$ applied to a multi-classifier system, as shown in

percentage change				
	Development			Evaluation
	total	D_1	D_2	
class A: imposter access	0.08%	0.14%	0.19%	0.01%
class B: genuine access	0.13%	0.25%	0.25%	0.04%
class $A+B$: all data	0.05%	0.09%	0.11%	0.01%

Table 5.2: Change in percentage results per one decision, Biosec database.

Figure 3.14. For the fingerprint modality, data samples originating from two separate fingerprint scanners are classified using the same matcher. Since for each finger the samples are collected using two different acquisition channels, they can be considered separate, mutually dependent modalities. Therefore the fingerprint example is used as an instance of $Q-stack$ applied to multimodal classification, according to Figure 3.15*. Finally, fusion of face and fusion modalities with the use of corresponding quality measures is considered. All available class sample sets, two for face and two for fingerprint, are used to evaluate the use of $Q-stack$ in a single base classifier scenario, according to the scheme shown in Figure 3.13. The details of the baseline classifiers, quality measures used in the experiments reported here are described in the respective Sections.

5.1.2 $Q-stack$ classifiers

As we have elicited in Chapter 3, the benefits of using class-independent quality measures in the classification process lies with the dependencies between the random variables that generate the components of the evidence vector. Therefore, a classifier that operates in the evidence space must aptly capture the nature of these dependencies. For instance, should a dependence relation between two evidence components be linear, a linear classifier that operates in the evidence space will be able to capture this dependence and consequently yield better classification results than a baseline classifier. On the other hand, non-linear inter-variate dependencies must be accounted for by a classifier capable of generating an non-linear decision hyper-surface in the evidence space.

In order to account for various possible relations between the evidence components, in the experiments reported here we utilize three different stacked classifiers, which we apply to classification in the evidence space according to the framework of $Q-stack$. These classifiers are:

Support Vector Machines with linear kernel (SVM-lin) The $SVM-lin$ classifier is a linear classifier that maximizes the classification margin between the classes in the same classification space where the evidence vectors live. Unlike the Linear Discriminants [42], the $SVM-lin$ does not make assumptions regarding the Gaussianity of the class-conditional joint distributions. Since it is a linear classifier in the evidence space, the $SVM-lin$ classifier is able to capture linear dependencies between the components of the evidence vector $\mathbf{e}$.

Support Vector Machines with radial basis functions kernel (SVM-rbf) The $SVM-rbf$ classifier utilizes the *kernel trick* and attempts to find a linear separating hyperplane in a transformed, arbitrarily high-dimensional space [152]. Projected back onto the original evidence space, the decision boundary generated by the $SVM-rbf$ classifier may be therefore nonlinear and of arbitrary complexity. The SVM-rbf classifier is capable of capturing non-linear dependencies between the evidence components.

*In the biometric literature, a fusion of classifiers operating on the same biometric modality (here: fingerprint) is often referred to as *intramodal fusion* [36].

Bayesian classifier using a GMM pdf estimator (Bayes) The complexity of the decision boundary of the Bayesian classifier is defined by the used approximations of the joint class-conditional distributions. In the experiments reported here we used Gaussian Mixture Models (GMM) [111] to approximate these distributions. The parameters of the GMM are estimated using the Expectation-Maximization (EM) algorithm, with the number of mixture components optimized on the development data set, with full covariance matrices.

The choice of these classifiers in the $Q-stack$ scenario was dictated by several rationales. First, the use of each of the classifiers puts different presumptions as to the inherent structure of the classified data. Namely, the use of the $SVM-lin$ classifier carries a tacit assumption that the classes can be linearly separated, the *Bayes* classifier (or rather its implementation used here) assumes that the processes that generated the data can be accurately approximated by linear combination of a preset and finite number of Gaussian distributions, and the $SVM-rbf$ classifier being the least restrictive in its assumptions. Second, they represent families of commonly used *discriminative* and *generative* classifiers [42, 111, 152], as we have discussed in Chapter 2.

5.1.3 Structure of the experiments and experimental reporting

The aim of the experiments reported in this chapter is to show that classification in the evidence space whose dimensions are defined by the components of evidence $\mathbf{e}$, as introduced in Chapter 3, yields lower classification errors if $\mathbf{e}$ contains classification scores and class-independent, score-dependent quality measures. At the same time, we show that addition of a class-independent quality feature that is independent on the remaining components of the evidence vector does not impact negatively the class separation. In turn, the improved class separation grants improved classification performance, provided that the chosen stacked classifier in the $Q-stack$ scheme is capable of aptly capturing the dependencies between the evidence components.

In the experiments reported here, we use two modalities, face and fingerprint.

Face In the case of the face modality, the very same face images are used for classification using two different base classifiers. We use two face quality measures that are characterized by different degrees of dependence on the scores originating from both baseline classifiers. We report results for $Q-stack$ for each of the face base classifiers separately, and for their fusion, with the use of proposed quality measures.

Fingerprint In the case of fingerprint modality, we use the same base classifier to classify two independently collected fingerprint impressions, originating from the same fingers but acquired using two different sensors. For the fingerprint experiments, we use two different fingerprint quality assessment algorithms, applied in an identical fashion to the fingerprint impressions coming from both acquisition channels. Consequently, we report results for $Q-stack$ for each of the acquisition channels separately, and for their multimodal fusion, using the available quality measures.

Face+Fingerprint fusion Finally, we report the results of $Q-stack$ applied to the fusion of face and fingerprint modalities. As baseline fusion results, we use all combinations of two baseline unimodal scores , where one of the scores originates from face, and the other from fingerprint modality. Then, we present results for a fusion of all four available unimodal baseline scores : two for face and two for fingerprint. The score-level fusion results are used as baseline for the $Q-stack$ experiments where the evidence vector is augmented by quality measures. Since in the fusion experiments we have all scores and all quality measures, and three stacked classifiers at our disposal, the number of possible $Q-stack$ experiments is very large. Reporting all

of the results for all combinations is beyond the frames of this thesis. We therefore limit the report to a selected choice of illustrative examples.

The experiments are conducted and reported in following order:

Marginal evidence distributions

For each of the baseline classifiers, we present the class-conditional score distributions, which gives the reader an understanding of the baseline class separation which the proposed method of $Q-stack$ is intended to improve using class-independent quality information. In order to show that the quality measures indeed do not convey usable class-selective information, we present the class-conditional distributions of all quality measures used in the experiments. To reinforce the graphical representation of the fact that for a given quality measure qm: $p(qm|A) \approx p(qm|B)$, we also present data classification results using only quality measures. The classification results close to chance-level of 0.5 confirm that the quality measures are largely class-independent. They therefore are individually irrelevant classification features and cannot be used for classification on their own. This determines that any improvement in classification due to the use of quality measures in the $Q-stack$ framework is due to the dependencies between the evidence components, and not to the class selectiveness of added quality measures.

Estimation of dependencies between evidence components

Since the quality measures are not class-selective, and therefore individually irrelevant to the classification task, we seek to explore their dependencies on the classification scores in order to afford improved class separation, and consequently more accurate classification. Since these dependencies are of crucial importance for the method proposed in this thesis, it is imperative to quantify the dependencies between the individual components of the evidence vectors, which later will be used in the classification process. These dependencies can be quantified in terms of mutual information between particular components of the evidence vector. Under a Gaussian assumption, it is easier to represent and intuitively interpret the inter-variate dependencies in terms of Pearson's correlation coefficients. However, the Gaussian assumption introduces estimation errors since most of the marginal class-conditional evidence distributions can only be crudely approximated by a Gaussian process, which the reader can appreciate from the graphical representations of these distributions for individual modalities. Therefore here we present the inter-variate dependency estimates for the components of the evidence vector in terms of both linear correlation coefficients and in terms of mutual information. In order to best understand the data structure we quote the inter-variate dependence estimates for each of the classes A and B individually, and for the entire development- and training data sets.

Given two random variables X and Y with given M instances, x and y, which are components of the evidence vector, the Pearson's correlation coefficient $\rho_{X,Y}$ between X and Y is given by [75]

$$\rho_{X,Y} = \frac{\Sigma_{i=1}^{M}(x-\mu_x)(y-\mu_y)}{(M-1)\sigma_x\sigma_y}, \tag{5.1}$$

where M is the sample size, μ_x and μ_y are means of x and y, and σ_x and σ_y are corresponding standard deviations of x and y. The mutual information $I_{X,Y}$ between X and Y is expressed by [22]

$$I_{X,Y} = \Sigma_X\Sigma_Y p(x,y) log(\frac{p(x,y)}{p(x)p(y)}). \tag{5.2}$$

In the experiments reported here we compute the estimates of correlation coefficient and of mutual information for the entire data set and for each class separately. We also provide a mean of

the dependence estimates computed individually for each class. The reason to do so is that the dependence estimates derived for each of the classes considered individually do not allow us to derive an estimate for the entire data set. As we have explained in Chapter 3, knowing the dependencies between the evidence components alone is not sufficient to predict the benefits coming from classification in the evidence space augmented by quality measures. However, together with a graphical representation of the joint class-conditional evidence distributions, the numerical dependence estimates allow for a better understanding of the results presented in this chapter.

Note that analyzing the differences in the corresponding inter-variate dependencies between the development and evaluation data sets help gauge how representative of the actual population both development and evaluation sample sets are. If these differ radically it can be concluded that either the evaluation and development sets are not drawn from the same population, or that at least one of them is not a representative sample. In either case, if the development sample is not representative of the population from which the evaluation sample is drawn, basic assumption of statistical pattern recognition is violated and obtained results have to be treated with caution. In the example considered here there exists a large discrepancy between the amounts of available data for each class, and much less data is available for development than for evaluation. Consequently, differences in corresponding inter-variate dependence estimates do exist. These differences and their consequences shall be discussed individually later in this chapter.

Classification in the evidence space

For each modality and their combination, available baseline scores and quality measures are combined into evidence vectors. Chosen combinations of evidence containing one or more baseline scores are classified according to the $Q-stack$ principles outlined in Chapter 3. In order to verify the claims worded in Section 5.1 set as objectives of this experimental work, baseline evidence combinations that do not include quality measures are compared to the corresponding combinations that include quality measures. In the experiments reported here we provide classification results on both development and evaluation data sets. The results obtained on the development data set are not a measure of how well does the proposed method perform - in principle as long as the set of evidence vectors belonging to class A is disjoint from the one belonging to class B, perfect class separation of the development can always be achieved. However, that would usually come at a cost of classifier overtraining. Therefore the actual performance of the $Q-stack$ technique is estimated based on the classification results obtained on the evaluation data set. The classification performance and class separation observed on the development set serve as an indicator of how well does the complexity of the decision boundary capture the dependencies between the evidence components. The interactions between the data- and classifier complexity can be best understood from their visual representations. For the cases where the $Q-stack$ classification is performed in a two-dimensional evidence space we provide a graphical representations of the joint class-conditional distributions and corresponding $Q-stack$ decision boundary in the evidence space.

As we have shown in Chapter 3, it is convenient to analyze the dependencies between particular components of the evidence vector in a pair-wise manner, graphically. As we have elicited previously, a departure of the optimal decision surface in the evidence space from that defined by the decision threshold obtained using the classification scores alone gives an idea of the magnitude of dependence between the scores and the considered quality measure. As a consequence of this dependence, the scale of the classification performance improvement due to the application of $Q-stack$ can be anticipated. Naturally, for higher-dimensional evidence spaces such representations are hardly feasible.

Results for all experiments are presented in terms of total classification accuracy AC, Half-

Total Error Rate $HTER$, where $HTER = \frac{1}{2}(ER_A + ER_B)$ [120], class errors: ER_A denoting the classification error per class A, and ER_B denoting the classification error per class B. Considering the fact that both the development and the evaluation data sets are unbalanced, i.e. there are far more data points in one class than in another, care must be taken when interpreting the presented results. In particular, one must bear in mind that the error resolution (the percentage change of error due to a change of class label assigned to one observation) is unequal for each class, yet when computing the total accuracy data points for both classes have the same impact. For this reason when analyzing both the total accuracy and the $HTER$, the latter one remains a more reliable performance estimator.

Comparative characteristics of performance for selected individual considered evidence combinations are also presented graphically in the form of DET curves in Appendix C.

Error prediction and rejection

In Chapter 4 we have described how the similarity score produced by a Bayesian classifier employed as a stacked classifier in the $Q - stack$ scenario can be considered as credences, or *degrees of belief* in the correctness of the classification decision. We have discussed how credence estimates can help predict and possibly reject misclassifications, thus boosting the factual accuracy of the classification system.

In this chapter we provide a practical embodiment of such credence estimators applied to the face and fingerprint modalities of the Biosec database. Since for both modalities, one of the considered stacked classifiers is a Bayes classifier, the matching scores it produces are credence estimates, as discussed in Chapter 4. Consequently, we use these credence estimates in order to predict misclassifications in both face and fingerprint modality. We compare the results of error prediction obtained using the classification scores alone with the error prediction results obtained using the scores and quality measures. We aim to demonstrate that the presented method of error prediction, as an functional extension of $Q - stack$, allows for accurate rejection of unreliable single classification decisions both in the case when the quality measures are available, and when they are not.

In the following sections we demonstrate the performance of the proposed error prediction and rejection techniques, applied to the classification decisions originating from single-classifier face and fingerprint matching, multi-classifier face matching, and to multimodal fingerprint matching. All error prediction experiments are conducted according to the procedure described in Chapter 4. First, baseline scores for one or both modalities are assembled together with considered quality measures into evidence vectors. Then, using the development set evidence vectors as classification features, a Bayesian stacked classifier is trained. So trained $Q - stack$ ensemble is then applied to classify the evaluation evidence vectors. The decision credence estimates produced by the stacked Bayesian classifier are then subjected to credence thresholding.

The credence estimates are evaluated in terms of the monotonicity and accountability criteria worded in Chapter 4. In order to verify if the monotonicity criterion is met we analyze the prediction performance of the credence estimates in terms of error reduction resulting from credence thresholding, and in terms of relative number of classification decision remaining after unreliable decisions have been discarded. In order to evaluate the error prediction performance in terms of the accountability criterion we report the difference between the mean credence estimates and the mean of the actual observed classification accuracy after 100 experimental runs. The necessity of using multiple experimental runs on the same data set is a consequence of the fact that used method of evidence modeling (EM) returns probability density models that depend heavily on the initialization parameters [15].

We provide experimental results of credence-based error rejection applied to multi-classifier sys-

tem of face verification and to multimodal fingerprint verification, but we do not extend the study to fusion of face and fingerprint. This restriction is dictated by two rationales. First, for the multimodal fusion the adopted parametric GMM-based distribution modeling approach turned out to be inadequate due to the insufficient amount of available data in respect to the high dimensionality of the multimodal evidence containing multiple quality measures. The credence estimates would then inherit the errors due to improper distribution modeling, which could result in meaningless error prediction results. Second, the fusion of face and fingerprint modalities resulted in extremely low error rates, which in turn means that there are very few errors to predict indeed. In such situation a proper evaluation of the accuracy of credence estimates was not feasible.

5.2 Face Modality

5.2.1 Baseline classifiers

The face images from the Biosec database contain the face itself surrounded by a substantial amount of non-uniform background. Therefore each face image is localized and normalized geometrically and photometrically prior to feature extraction and matching. In practical applications the task of face detection and localization in the original image is delegated to a dedicated classifier (face detector)[99]. As such, face detectors are known to make mistakes, in particular in images of low image quality. The mistakes of automatic face detector are bound to have a negative impact not only on the face matching performance afterwards [106, 137], but also on the values of extracted quality measures, since the quality measures are taken from an already localized and normalized image [73]. Therefore in the experiments reported in this thesis we decided to rule out the impact of potentially inaccurate face detection algorithms and consequently to work with faces manually located in the original images. All images were normalized geometrically by aligning the positions of the eye centers. The images were normalized photometrically using the anisotropic diffusion-based method proposed in [64].

The specific experimental protocol defined for the Biosec database results in certain important limitations regarding the construction of the baseline classifier. Namely, since only one training sample per client for each successive comparison is at hand, it is not possible to derive an estimate of the client-specific genuine access score distribution. Furthermore, classical subspace projection-based discriminative methods are not applicable [13], since no clients whose images appear in the evaluation set are represented also in the development set. For this reason we have decided to use two classifiers, a Bayesian classifier using a model adaptation scheme [126], and a PCA-based approach [145, 157]. The details of the used face classifiers are as follows:

1. **DCT** - local $DCTmod2$ features and a Bayes classifier based on the feature distributions approximated by Gaussian Mixture Models (GMM)[137]. The classifier uses two concurrent GMM models, one for class A (imposter attempt) and another for class B (genuine access). While the principles of the classifier training and model construction follow [137], the fact that the experimental protocol assumes one-to-one matching, the implementation of the classifier is unusual and deserves a more detailed explanation. During the classifier training phase, a generic GMM face model is built using the $DCTmod2$ features extracted from all face images from the training data set. The model is referred to as the *world model*. In the testing phase, one single training image is not enough to train a full client model and therefore a Maximum-A-Posteriori (MAP) adaptation [126] of the world model is applied in order to construct a client model for a particular matching. During the testing phase, posterior probability of the genuine access class A given the $DCTmod2$ features extracted from a testing image is

computed. The value of this probability is compared to a baseline decision threshold τ_{f1} estimated on the development set. To the best of our knowledge, this is the first report on face matching with MAP-adapted models, where only one single sample is used for client-specific model adaptation.

2. **PCA** - Mahalanobis distance between global PCA feature vectors [157]. The PCA projection space was found using all images from the development data set. During the testing phase both matched images are projected onto the subspace defined by eigenvectors associated with a set of 40 largest eigenvalues. The similarity score was found by computing the Mahalanobis distance between both projections. The covariance matrix that describes the dependencies and scaling of individual projected dimensions was found by analyzing projections of all images from the development set onto the new space of reduced dimensionality. In order to arrive at a classification decision a score threshold τ_{f2} is applied. The value of τ_{f2} is estimated on the development data set, part D_2.

The two face matchers were chosen for the reported experiments since they both operate on very different features. The local $DCTmod2$ features encode mostly high spacial frequencies, while the projection of the face images on the PCA subspace emphasizes lower spacial frequencies. The scores produced by the DCT matcher are denoted as x_{f1} and the scores produced by the PCA matcher are denoted as x_{f2}. We stress here that since the main topic of this thesis is improving classification with quality measures, we do not claim to have constructed optimal, best possible classifier for the Biosec database. Although we have given every effort to ensure that the parameters of the constructed baseline classifiers are tuned to maximize classification performance, we assume that another feature extraction or classification algorithm may be able to deliver lower baseline classification errors. This, however, does not impair the soundness of the findings presented.

5.2.2 Quality measures

According to the theoretical developments presented in Chapter 3, the quality measures must be dependent on the similarity scores of the classifier they operate with in order to help separate between classes. A properly designed quality measure must therefore be able to capture those characteristic features of classified signals that are also of importance from the viewpoint of the classification process. In the case of face recognition, the choice of proper quality measures is not an obvious task since it is by itself not obvious which face features decide about a successful facial recognition [6, 59].

Rather than trying to solve this problem, we approach the issue of constructing a relevant quality measure by looking at the actual features that the designed baseline classifiers are using. We discuss here two face image quality measures - a normalized two-dimensional cross-correlation coefficient with an average face template, and a probabilistic quality measure based on the DCT classification features. The same quality measures, which we first introduced in [87, 88], are used in the experiments reported in this section. The first of these quality measures is by design related to the DCT classifier, while the latter is related to the PCA classifier. Below follow the details of the used quality measures.

Sum of likelihoods

In this section we provide an example of quantifying the quality of face image based on the classification features. The presented quality measure is designed to work with the DCT face classifier.

The *DCT* classifier is essentially a Bayes classifier that takes the classification decision for observed classification features f based on log-likelihood ratio:

$$\begin{aligned} \forall_f : log(p(f|A)) - log(p(f|B)) \geq 0 \Rightarrow decide : A \\ \forall_f : log(p(f|A)) - log(p(f|B)) < 0 \Rightarrow decide : B, \end{aligned} \tag{5.3}$$

which is a numerically efficient variant of the likelihood ratio-based decision rule, assuming equal prior probabilities $P(A)$ and $P(B)$. The likelihood ratio-based decision rule, as a direct consequence of the Neyman-Person's theorem [42], has been shown to be optimal [12] for fixed-length vectors f.

Obviously, in order to guarantee best classifier performance on clean, high-quality data we wish to create models that best account for this data, automatically leaving out lower-quality samples. Otherwise, accounting for low quality data in the models is bound to compromise the classification performance on high-quality data [85, 86]. Therefore in order to be able to treat the low-quality samples appropriately without sacrificing good performance on high-quality samples, it is necessary to somehow quantify the sample quality. Since by definition the models account for good quality data, the samples that the models do not well account for can be considered of degraded quality. Note, however, that the magnitude of the likelihood ratio does not at all deliver any information as to how well do the models account for the observed features extracted from the given sample.

The likelihood ratio merely tells *how many times* is the value of the class-conditional distribution $p(f|B)$ smaller or greater than $p(f|A)$ at given observed f. It is enough to make the optimal decision given observed f, but in fact both $p(f|B)$ and $p(f|A)$ can be arbitrarily small for any computed value of the likelihood ratio. If both $p(f|B)$ and $p(f|A)$ are small, that hints that the observed feature vector f is well accounted for by neither $p(f|B)$ nor $p(f|A)$, and consequently can be considered to be of *low quality*, where quality is regarded in comparative terms. This reasoning is further reinforced by the actual interpretation of the models $p(f|B)$ and $p(f|A)$. Namely, routinely in biometric identity verification one of them is called *world model* and the other *client model*. The *world model* is supposed to be encoding the average traits of the population, while the *client model* emphasizes particular characteristic traits of an individual that makes him stand out of the population. For a genuine client access, therefore (class B in our experiments), it is expected that the *client model* will "explain" the observed features better than the *world model*. For an imposter attach, the *world model* should "explain" the observed features better since it is more likely that the imposter's data will share similar features with the entire population, rather than with the particular person he is trying to impersonate. In other words, for the samples of nominal quality, either of the models should account well for the observed data. Therefore, if none of them actually does, the sample can be considered as low quality.

This reasoning can be formalized as follows. The value of $p(f|A)$ and $p(f|B)$ at the newly observed feature vector f_0 determine how well does each of the models account for the observed f_0. We demand that f_0 is well-accounted for by either of them, therefore

$$qm = p(f_0|A) + p(f_0|B). \tag{5.4}$$

note that the quality measure expressed by Equation 5.4 is defined in feature rather than in signal domain. It therefore takes into account strictly the same information as is used during for classification. For this reason a close coupling and dependence between the quality measure and the classification scores can be expected.

During the matching process two images are compared and as a result a single scalar, the matching similarity score, is returned. At the same time, the correlation with the quality measure is computed for each of the matched images individually. It is therefore necessary to arrive at one

quality measure that represents the quality of the matched sample pair. If qm_0 and qm_1 are the quality measures computed for each of the matched face images then the resulting combined quality measure $qm_{0,1}$ used as evidence is computed, following [48], as

$$qm_{0,1} = \sqrt{qm_0 qm_1}. \tag{5.5}$$

In the experiments reported here this quality measure is referred to as qm_{f1}

2D cross-correlation coefficient with an average face template

In order to properly explain how to construct the quality measure that will work with the PCA classifier we first provide a recap on the features that the PCA classifier uses. The PCA classifier is based on the idea that face images can be represented as points in m-dimensional space, where $m = l_x \times l_y$, and l_x and l_y are the dimensions of the face image measured in pixels. Each of the coordinates of this point corresponds to one particular pixel location in the face image and takes the value of its pixel intensity. Since the number of dimensions, depending on the size of the used face images, can become very large, classification may consequently become difficult due to the *dimensionality curse*. The main idea behind the PCA feature extraction is to reduce the dimensionality of the classification problem by projecting the points that live in the m-dimensional space onto a lower-dimensional subspace. the projection is chosen in such a way that it preserves the most significant variability of the data, which can be considered a form of image compression. In the experiments reported here we adopted a method proposed by Sirovich and Kirby in [145] and popularized by the influential paper by Turk and Pentland [157], where the projection is found by performing Principal Component Analysis on a grayscale pixel value covariance matrix of development images, reshaped into vectors.

Let $\Gamma_1, \Gamma_2, ..., \Gamma_M$ be vectors created by reshaping M development images, and let $\Psi = \frac{1}{M}\sum_{i=1}^{M}\Gamma_i$ and $\Phi_i = \Gamma_i - \Psi$. The optimal projection of the original images onto a new space is found by computing M eigenvectors u_i and associated eigenvalues λ_i of a covariance matrix C:

$$C = \frac{1}{M}\sum_{j=1}^{M}\Phi_j\Phi_j^T = AA^T, \tag{5.6}$$

where matrix $A = [\Phi_1\ \Phi_2\ ...\ \Phi_M]$. Only k first eigenvectors corresponding to k largest eigenvalues are in fact useful in representing faces, and a low dimensional representation ω_i of a given face Γ_i is found from

$$\omega_i = u_k^T(\Gamma_i - \Psi). \tag{5.7}$$

In our experiments we used $k = 40$ eigenvectors, following [157, 158]. The classification similarity score between two face images Γ_a and Γ_b is found by computing Mahalanobis distance between their projections, ω_a and ω_b:

$$x_{f2}(\Gamma_a, \Gamma_b) = \sqrt{(\omega_a - \omega_b)\Sigma_x(\omega_a - \omega_b)}, \tag{5.8}$$

where the $k \times k$-element covariance matrix Σ_x is estimated using the development set.

We construct an average face template as follows. Given the set of development images Γ_1, Γ_2, ..., Γ_M we compute their low-dimensional projection onto the space defined by chosen k eigenvectors according to Equation 5.7, obtaining $\omega_1, \omega_2, ..., \omega_M$. An average face representation in the

transformed space is given by:

$$\bar{\omega} = \frac{1}{M}\sum_{j=1}^{M}\omega_j. \tag{5.9}$$

And the inverse transformation that brings the average face back into an average face template $\bar{\Gamma}$ the original face space is given by

$$\bar{\Gamma} = \bar{\omega}u_k^T + \Psi. \tag{5.10}$$

The obtained average face template $\bar{\Gamma}$, after reordering the pixel values back onto a rectangular matrix of original size of $l_x \times l_y$, is an average face reconstruction from only selected eigenvalues $\omega_{1,2,\ldots,k}$ and therefore contains only the strongest and most prominent averaged features of the faces from the development set. The PCA transformation serves this purpose well since is produces a projection into a subspace where the original signals are best represented. The average face template used in the experiments reported in this chapter is shown in Figure 5.2.

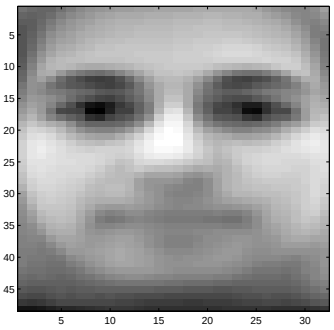


Figure 5.2: Average face template $\bar{\Gamma}$ used in face image quality estimation.

It is apparent that high-resolution details are lost during the reconstruction process, while low-frequency features, such as head pose and illumination, are preserved. The reader can agree that the face image shown in 5.2 indeed encodes the major facial features and their spatial interrelations, while it hardly captures any individual traits. Therefore comparing an actual face image Γ_0 to the average face template gives a measure of how much does Γ_0 resemble the quality of the set of development face images, again in terms of head pose, illumination, and geometrical alignment and all features that can be visually appreciated when inspecting $\bar{\Gamma}$. The measure of similarity between Γ_0 and $\bar{\Gamma}$ is computed as follows:

$$qm_0 = \frac{\Sigma_{i=1}^{l_x l_y - 1}(p_{\Gamma_0} - \mu_{\Gamma_0})(p_{\bar{\Gamma}} - \mu_{\bar{\Gamma}})}{(l_x l_y - 1)\sigma_{\Gamma_0}\sigma_{\bar{\Gamma}}}, \tag{5.11}$$

where μ_{Γ_0} and σ_{Γ_0} are mean and variance of all pixels in Γ_0, and $\mu_{\bar{\Gamma}}$ and $\sigma_{\bar{\Gamma}}$ are mean and variance of all pixels in $\bar{\Gamma}$. The quality measure qm_0 is Pearson's correlation coefficient between Γ_0 and $\bar{\Gamma}$.

As it was the case with the quality measure discussed in Section 5.2.2, for a compared pair of face images we arrive at two quality measures, each computed according to Equation 5.11. A composite quality measure for the comparison is then computed using Equation 5.5. In the experiments reported here this quality measure is referred to as qm_{f2}. The proposed quality measure qm_{f2} is

closely linked with the nature of the PCA classifier since the average face template is built using face image projections onto the very same space in which the classification features live. Therefore these aspects of human face appearance that are captured by the eigenvectors u_k are also captured by the average face template.

5.2.3 $Q-stack$ in face verification

In order to verify the claim that an inclusion of quality measures in the evidence vector allows for more accurate classification in the $Q-stack$ framework than using baseline classifiers, we conducted a series of experiments with various configurations of available evidence. In particular, the experiments aimed at showing that those configurations of evidence which include quality measures as classification features to the stacked classifier give better classification results than baseline systems that use no quality measures.

The experimental results for face modality are reported according to the order spelled out in Section 5.1.3. The available evidence components for the face modality include:

DCT **baseline classifier:** Similarity scores x_{f1}, Quality measures qm_{f1}

PCA **baseline classifier:** Similarity scores x_{f2}, Quality measures qm_{f2}

In the following section we provide a comparison of baseline classification results obtained for the face modality, with the results obtained with the use of quality measures according to the proposed method of $Q-stack$. The baseline results are obtained for:

Single-classifier system Two single-classifier systems are considered that use matching scores x_{f1} and x_{f2}, respectively.

Multiple-classifier system A multiple-classifier system is considered that uses the scores from both baseline classifiers, $\mathbf{e} = [x_{f1}, x_{f2}]$.

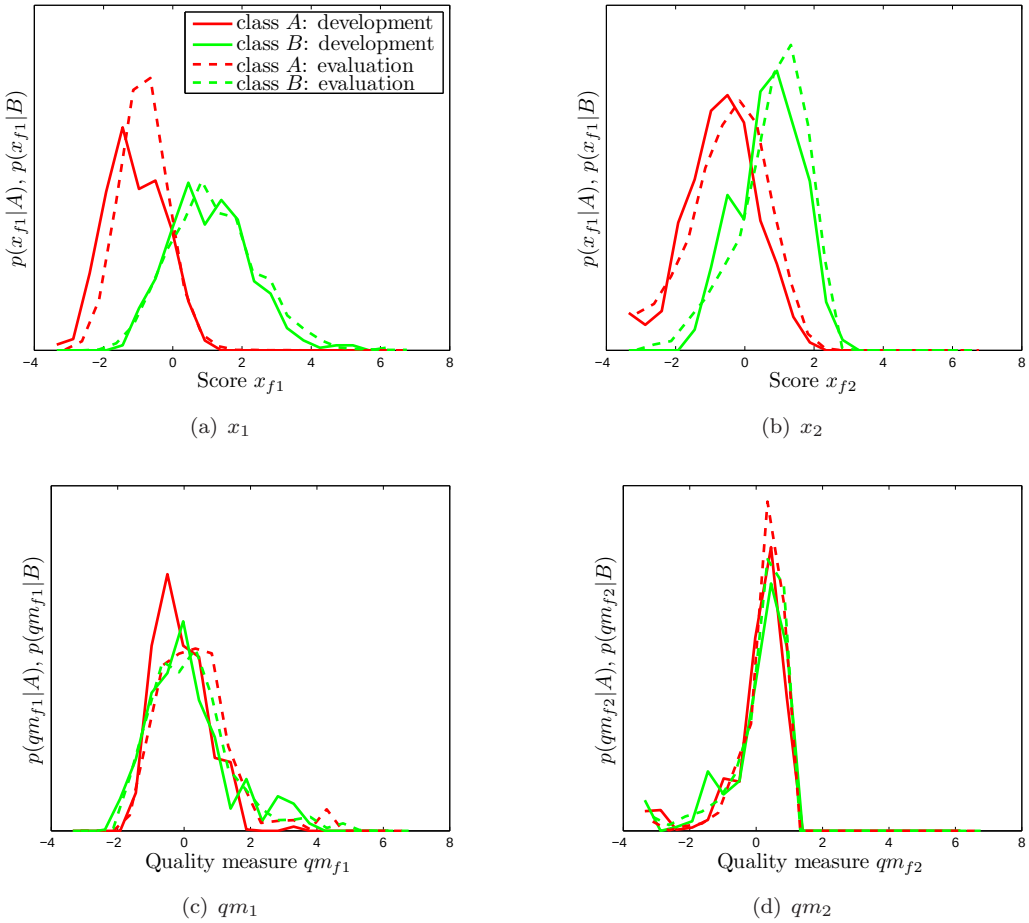
Analysis of dependencies between evidence components

Figure 5.3 shows the distribution of all evidence components available for the face modality:

- Classification similarity scores for the DCT classifier, x_{f1}
- Classification similarity scores for the PCA classifier, x_{f2}
- Sum of likelihoods, a quality measure designed for the DCT classifier, qm_{f1}
- Correlation coefficient with the average face template, a quality measure designed for the PCA classifier, qm_{f2}

In order to show that the quality measures indeed do not allow for a meaningful classification of the signals without the use of the classification scores, Table 5.3 presents the corresponding classification results on both development and evaluation data sets. As it is evident from Figures 5.3 (c),(d) and from Table 5.3, quality measures alone are individually irrelevant features and therefore of little use for classification by themselves.

We therefore make use of quality measures in ensemble with the classification scores, in whose context the quality information becomes conditionally relevant to the classification task. As theoretically shown in chapter 3, in order to improve class separation in the evidence space, where the evidence consists of class-dependent and class-independent components, pair-wise relations of



(a) x_1 (b) x_2

(c) qm_1 (d) qm_2

Figure 5.3: Class-conditional evidence distributions $p(e|A)$ and $p(e|B)$.

	AC	$HTER$	ER_A	ER_B
development				
$\mathbf{e} = [qm_{f1}]$	0.561	0.438	0.435	0.441
$\mathbf{e} = [qm_{f2}]$	0.541	0.443	0.388	0.499
evaluation				
$\mathbf{e} = [qm_{f1}]$	0.566	0.464	0.510	0.418
$\mathbf{e} = [qm_{f2}]$	0.536	0.446	0.418	0.474

Table 5.3: Classification using only quality measures, face, Biosec database.

development					evaluation				
A									
	x_{f1}	x_{f2}	qm_{f1}	qm_{f2}		x_{f1}	x_{f2}	qm_{f1}	qm_{f2}
x_{f1}	1.00	0.35	0.26	0.26	x_{f1}	1.00	0.23	-0.15	0.03
x_{f2}	0.35	1.00	-0.10	0.62	x_{f2}	0.23	1.00	-0.60	0.63
qm_{f1}	0.26	-0.10	1.00	-0.23	qm_{f1}	-0.15	-0.60	1.00	-0.69
qm_{f2}	0.26	0.62	-0.23	1.00	qm_{f2}	0.03	0.63	-0.69	1.00
B									
	x_{f1}	x_{f2}	qm_{f1}	qm_{f2}		x_{f1}	x_{f2}	qm_{f1}	qm_{f2}
x_{f1}	1.00	0.14	0.37	-0.26	x_{f1}	1.00	0.15	0.31	-0.40
x_{f2}	0.14	1.00	-0.38	0.54	x_{f2}	0.15	1.00	-0.44	0.53
qm_{f1}	0.37	-0.38	1.00	-0.71	qm_{f1}	0.31	-0.44	1.00	-0.71
qm_{f2}	-0.26	0.54	-0.71	1.00	qm_{f2}	-0.40	0.53	-0.71	1.00
$A+B$									
	x_{f1}	x_{f2}	qm_{f1}	qm_{f2}		x_{f1}	x_{f2}	qm_{f1}	qm_{f2}
x_{f1}	1.00	0.55	0.29	-0.01	x_{f1}	1.00	0.36	0.18	-0.11
x_{f2}	0.55	1.00	-0.12	0.47	x_{f2}	0.36	1.00	-0.39	0.50
qm_{f1}	0.29	-0.12	1.00	-0.47	qm_{f1}	0.18	-0.39	1.00	-0.60
qm_{f2}	-0.01	0.47	-0.47	1.00	qm_{f2}	-0.11	0.50	-0.60	1.00
$\overline{(A,B)}$									
	x_{f1}	x_{f2}	qm_{f1}	qm_{f2}		x_{f1}	x_{f2}	qm_{f1}	qm_{f2}
x_{f1}	1.00	0.25	0.32	0.00	x_{f1}	1.00	0.19	0.08	-0.19
x_{f2}	0.25	1.00	-0.24	0.58	x_{f2}	0.19	1.00	-0.52	0.58
qm_{f1}	0.32	-0.24	1.00	-0.47	qm_{f1}	0.08	-0.52	1.00	-0.70
qm_{f2}	0.00	0.58	-0.47	1.00	qm_{f2}	-0.19	0.58	-0.70	1.00

Table 5.4: Pearson's pair-wise correlation coefficients ρ between components of the evidence vector, face modality, Biosec database.

dependence must exist between the evidence components. In Tables 5.4 and 5.5 we present the estimates of these pair-wise dependencies between scores x_{f1}, x_{f2} and quality measures qm_{f1} and qm_{f2}, in terms of Person's correlation coefficient ρ (Equation 5.1) and in terms of mutual information I (Equation 5.2). The results are reported for each class individually (A,B), for the entire data set ($A+B$), and as an average of the dependence estimates for each class, $\overline{(A,B)}$.

As the reader can gather from Tables 5.4 and 5.5, the general tendencies in these estimates appear fairly consistent, and show a strong dependence between the scores x_{f1} and x_{f2}, which was to be expected since both classifiers operate on the same modality. Further, the dependence between x_{f2} and qm_{f2} is well-pronounced, and a somewhat weaker but still clear dependence between x_{f1} and qm_{f1} is evident. Therefore the dependencies between the quality measures and the classification scores are expected to deliver improved separation between classes A and B in the evidence space in respect to the class separation using marginal score distributions alone.

Classification in the evidence space

In this section we show how the decision boundaries set by all three considered stacked classifiers listed in section 5.1.2 differ from the baseline classification thresholds $x = \tau_{f1}$ and $x = \tau_{f2}$. For clarity reasons here we limit the visualization of the $Q-stack$ classification boundaries to the cases where

development					evaluation				
A									
	x_{f1}	x_{f3}	qm_1	qm_2		x_{f1}	x_{f3}	qm_1	qm_2
x_{f1}	-	0.11	0.08	0.10	x_{f1}	-	0.09	0.11	0.06
x_{f3}	0.11	-	0.04	0.35	x_{f3}	0.09	-	0.30	0.27
qm_1	0.08	0.04	-	0.10	qm_1	0.11	0.30	-	0.39
qm_2	0.10	0.35	0.10	-	qm_2	0.06	0.27	0.39	-
B									
	x_{f1}	x_{f3}	qm_1	qm_2		x_{f1}	x_{f3}	qm_1	qm_2
x_{f1}	-	0.08	0.28	0.19	x_{f1}	-	0.15	0.26	0.22
x_{f3}	0.08	-	0.20	0.29	x_{f3}	0.15	-	0.36	0.33
qm_1	0.28	0.20	-	0.46	qm_1	0.26	0.36	-	0.61
qm_2	0.19	0.29	0.46	-	qm_2	0.22	0.33	0.61	-
$A+B$									
	x_{f1}	x_{f3}	qm_1	qm_2		x_{f1}	x_{f3}	qm_1	qm_2
x_{f1}	-	0.30	0.20	0.17	x_{f1}	-	0.15	0.10	0.04
x_{f3}	0.30	-	0.09	0.29	x_{f3}	0.15	-	0.12	0.21
qm_1	0.20	0.09	-	0.24	qm_1	0.10	0.12	-	0.25
qm_2	0.17	0.29	0.24	-	qm_2	0.04	0.21	0.25	-
$\overline{(A,B)}$									
	x_{f1}	x_{f3}	qm_1	qm_2		x_{f1}	x_{f3}	qm_1	qm_2
x_{f1}	-	0.09	0.18	0.15	x_{f1}	-	0.12	0.18	0.14
x_{f3}	0.09	-	0.12	0.32	x_{f3}	0.12	-	0.33	0.30
qm_1	0.18	0.12	-	0.28	qm_1	0.18	0.33	-	0.50
qm_2	0.15	0.32	0.28	-	qm_2	0.14	0.30	0.50	-

Table 5.5: Mutual information estimates between components of the evidence vector, face modality, Biosec database.

the evidence vector contains only two components. In all Figures the decision boundaries obtained by particular stacked classifiers are consistently shown in different colors, namely: $SVM - lin$ - black; $SVM - rbf$ - magenta; $Bayes$ - blue. In each of the presented figures corresponding baseline score thresholds are marked by dashed lines.

Figures 5.18, 5.16 and 5.17 show classification in the evidence space $\mathbf{e} = [x_{f1}, x_{f2}]$, which is equivalent to trained multiclassifier fusion of scores x_{f1} and x_{f2}. In these figures both baseline thresholds are shown simultaneously. The results of classification in the evidence space defined by $\mathbf{e} = [x_{f1}, x_{f2}]$ is considered baseline for the $Q - stack$ systems applied to multi-classifier fusion with quality measures.

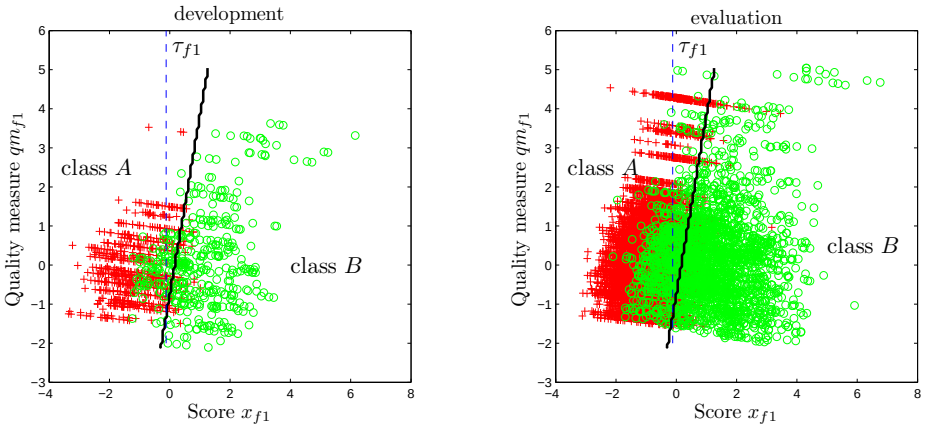


Figure 5.4: Classification in the evidence space using the $SVM - lin$ stacked classifier, face modality, joint evidence distributions $\mathbf{e} = [x_{f1}, qm_{f1}]$.

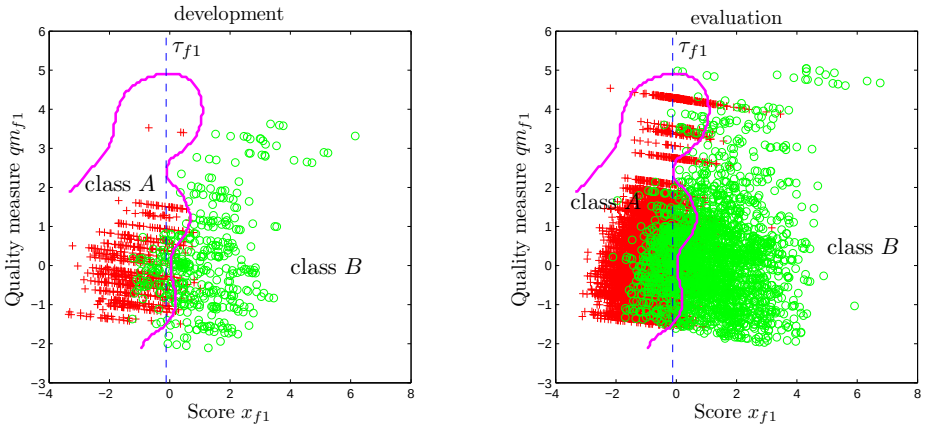


Figure 5.5: Classification in the evidence space using the $SVM - rbf$ stacked classifier, face modality, joint evidence distributions $\mathbf{e} = [x_{f1}, qm_{f1}]$.

Table 5.6 gives the classification results for the face modality in terms of total accuracy AC, $HTER$, and error rates per class ER_A and ER_B, for the development data set. Table 5.7 gives the classification results in terms of total accuracy, $HTER$, and error rates per class ER_A and ER_B, for the evaluation data set. In both Tables 5.6 and 5.7 the classification results for all three stacked classifiers, $SVM - lin$, $SVM - rbf$ and $Bayes$, are listed.

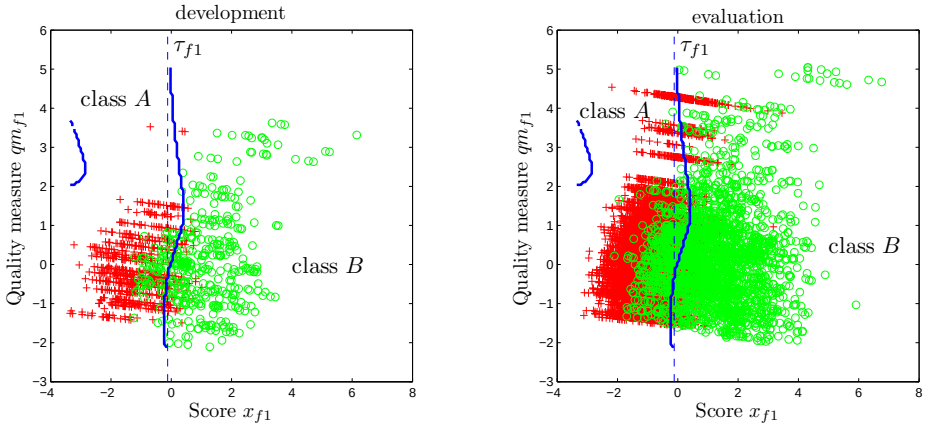


Figure 5.6: Classification in the evidence space using the *Bayes* stacked classifier, face modality, joint evidence distributions $\mathbf{e} = [x_{f1}, qm_{f1}]$.

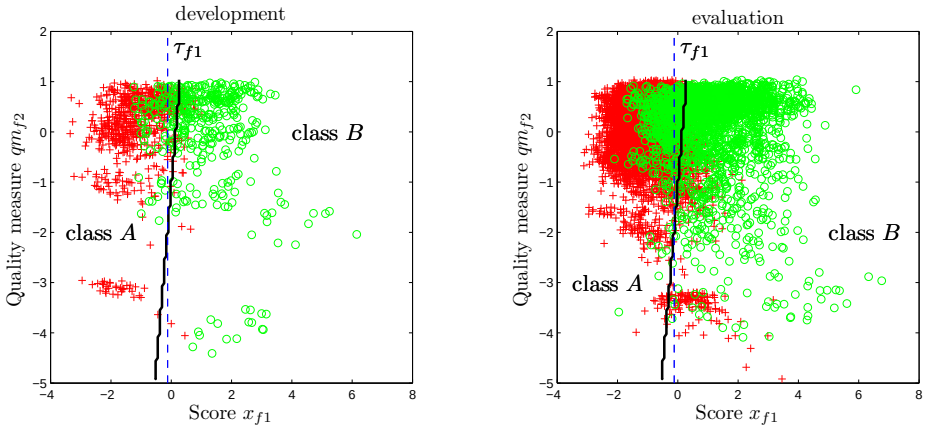


Figure 5.7: Classification in the evidence space using the $SVM - lin$ stacked classifier, face modality, joint evidence distributions $\mathbf{e} = [x_{f1}, qm_{f2}]$.

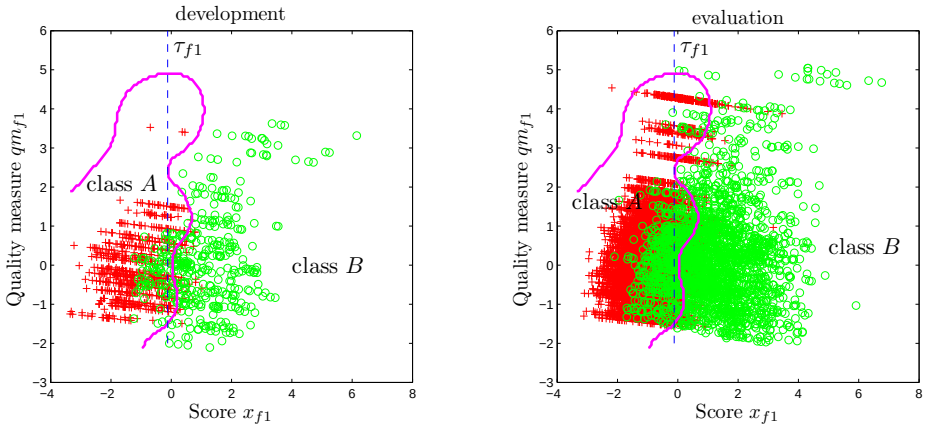


Figure 5.8: Classification in the evidence space using the $SVM - rbf$ stacked classifier, face modality, joint evidence distributions $\mathbf{e} = [x_{f1}, qm_{f2}]$.

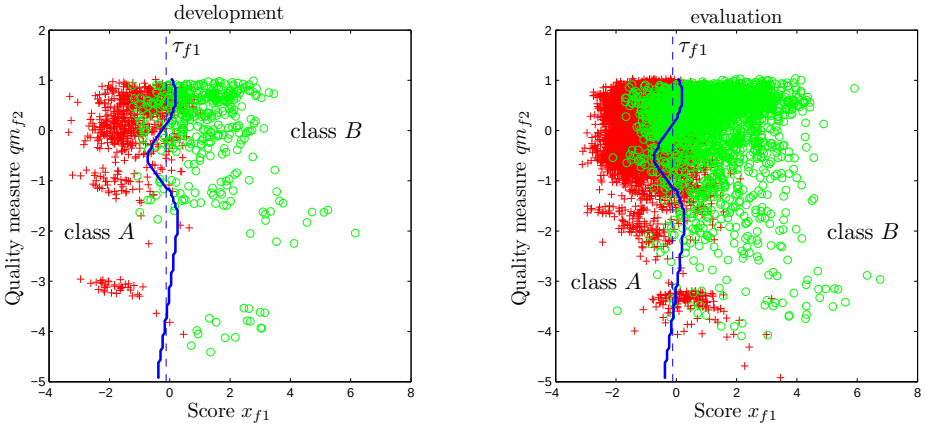


Figure 5.9: Classification in the evidence space using the *Bayes* stacked classifier, face modality, joint evidence distributions $\mathbf{e} = [x_{f1}, qm_{f2}]$.

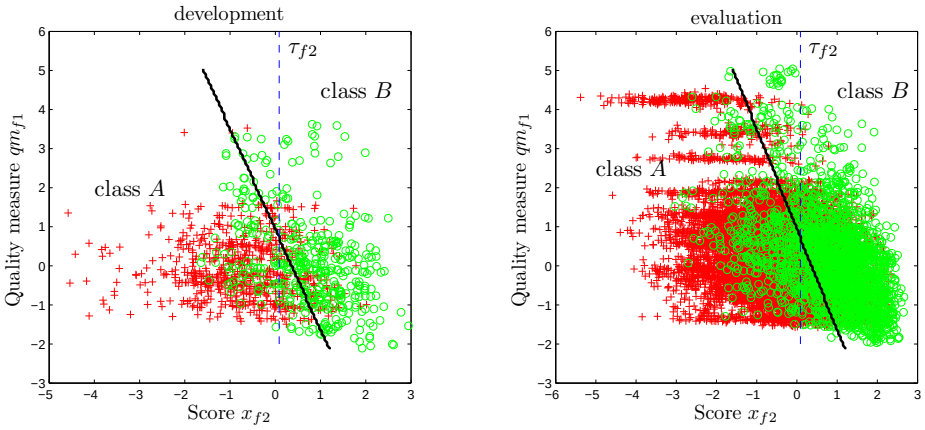


Figure 5.10: Classification in the evidence space using the $SVM - lin$ stacked classifier, face modality, joint evidence distributions $\mathbf{e} = [x_{f2}, qm_{f1}]$.

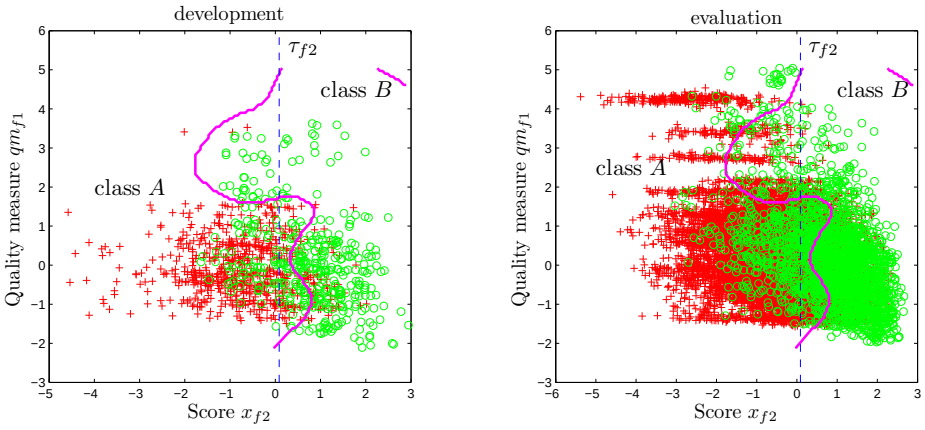


Figure 5.11: Classification in the evidence space using the $SVM - rbf$ stacked classifier, face modality, joint evidence distributions $\mathbf{e} = [x_{f2}, qm_{f1}]$.

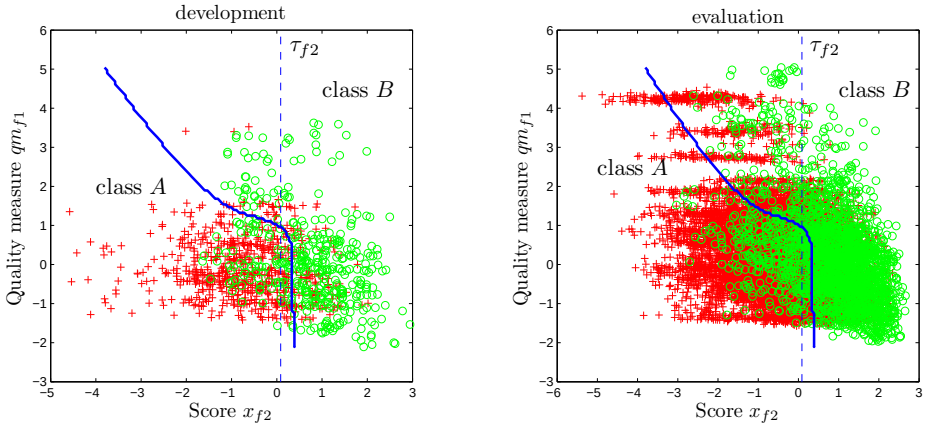


Figure 5.12: Classification in the evidence space using the *Bayes* stacked classifier, face modality, joint evidence distributions $\mathbf{e} = [x_{f2}, qm_{f1}]$.

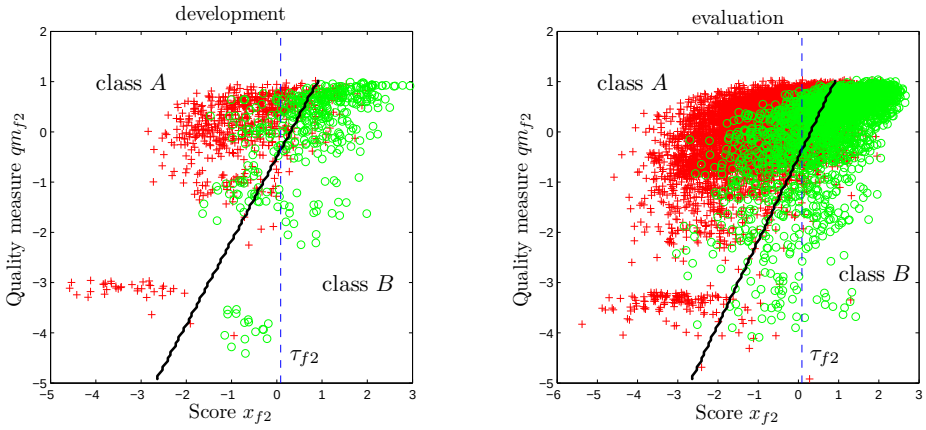


Figure 5.13: Classification in the evidence space using the $SVM - lin$ stacked classifier, face modality, joint evidence distributions $\mathbf{e} = [x_{f2}, qm_{f2}]$.

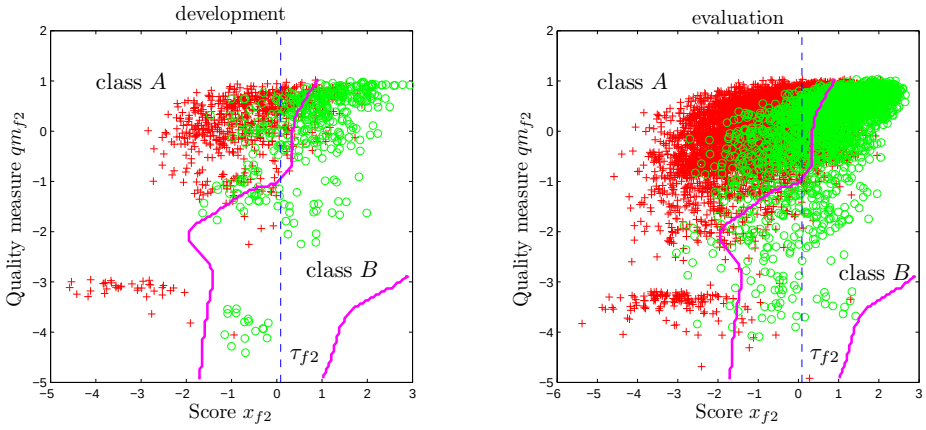


Figure 5.14: Classification in the evidence space using the $SVM - rbf$ stacked classifier, face modality, joint evidence distributions $\mathbf{e} = [x_{f2}, qm_{f2}]$.

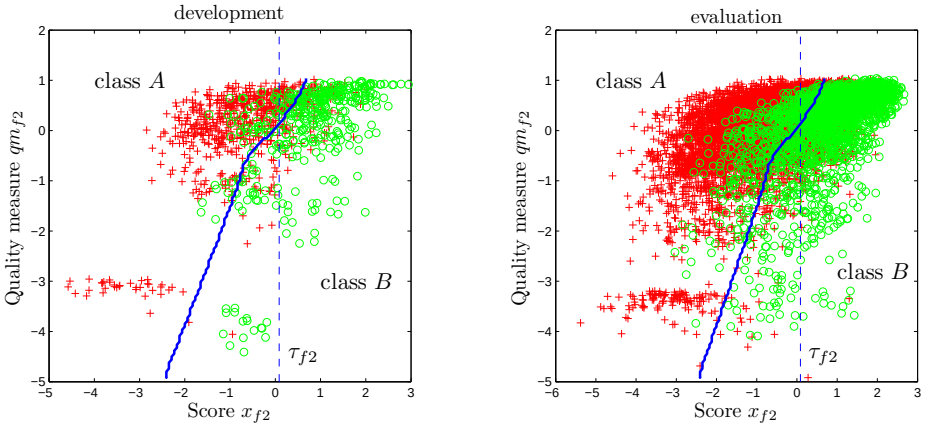


Figure 5.15: Classification in the evidence space using the *Bayes* stacked classifier, face modality, joint evidence distributions $\mathbf{e} = [x_{f2}, qm_{f2}]$.

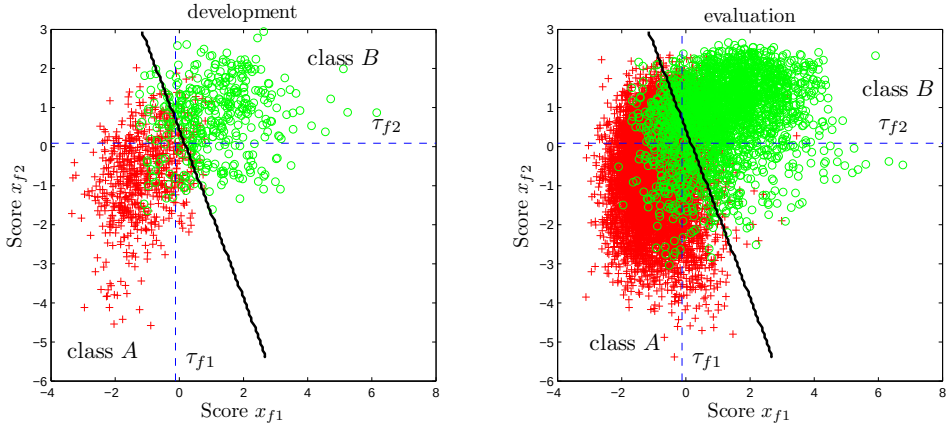


Figure 5.16: Classification in the evidence space using the $SVM - lin$ stacked classifier, face modality, joint evidence distributions $\mathbf{e} = [x_{f1}, x_{f2}]$.

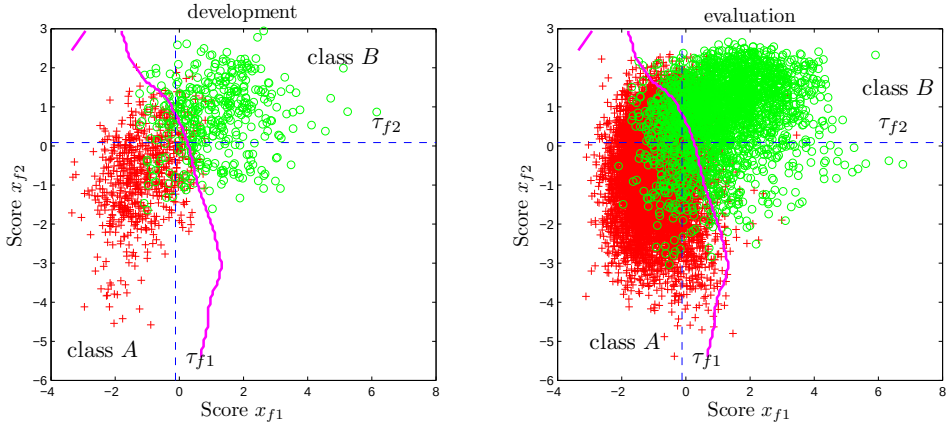


Figure 5.17: Classification in the evidence space using the $SVM - rbf$ stacked classifier, face modality, joint evidence distributions $\mathbf{e} = [x_{f1}, x_{f2}]$.

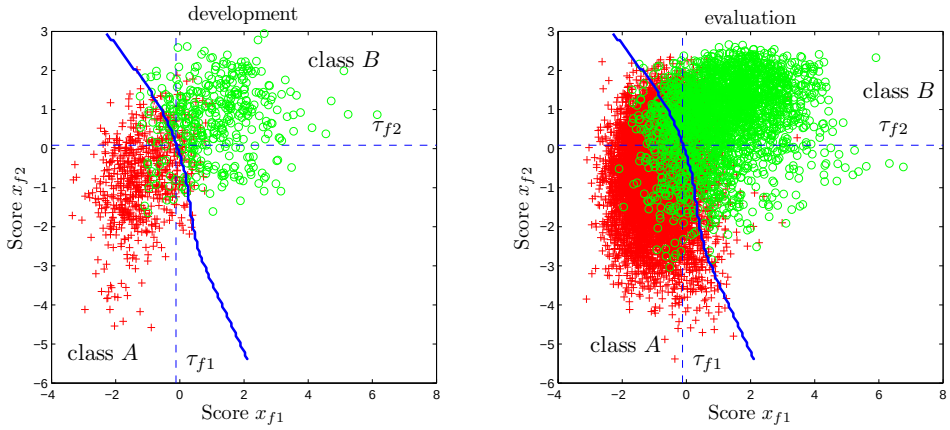


Figure 5.18: Classification in the evidence space using the *Bayes* stacked classifier, face modality, joint evidence distributions $\mathbf{e} = [x_{f1}, x_{f2}]$.

	AC	$HTER$	ER_A	ER_B
baseline				
$\mathbf{e} = x_{f1}$	*0.859*	*0.142*	*0.143*	*0.141*
$\mathbf{e} = x_{f2}$	*0.764*	*0.239*	*0.250*	*0.228*
$SVM - lin$				
$\mathbf{e} = [x_{f1}, qm_{f1}]$	0.857	0.140	0.130	0.151
$\mathbf{e} = [x_{f1}, qm_{f2}]$	0.870	0.136	0.160	0.112
$\mathbf{e} = [x_{f1}, qm_{f1}, qm_{f2}]$	0.871	0.130	0.133	0.127
$\mathbf{e} = [x_{f2}, qm_{f1}]$	0.785	0.218	0.228	0.208
$\mathbf{e} = [x_{f2}, qm_{f2}]$	0.809	0.193	0.198	0.188
$\mathbf{e} = [x_{f2}, qm_{f1}, qm_{f2}]$	0.805	0.192	0.183	0.202
$\mathbf{e} = [x_{f1}, x_{f2}]$	*0.887*	*0.113*	*0.110*	*0.115*
$\mathbf{e} = [x_{f1}, x_{f2}, qm_{f1}]$	0.889	0.114	0.123	0.105
$\mathbf{e} = [x_{f1}, x_{f2}, qm_{f2}]$	0.903	0.100	0.110	0.090
$\mathbf{e} = [x_{f1}, x_{f2}, qm_{f1}, qm_{f2}]$	0.898	0.098	0.083	0.114
$Bayes$				
$\mathbf{e} = [x_{f1}, qm_{f1}]$	0.861	0.140	0.145	0.135
$\mathbf{e} = [x_{f1}, qm_{f2}]$	0.862	0.141	0.153	0.129
$\mathbf{e} = [x_{f1}, qm_{f1}, qm_{f2}]$	0.860	0.141	0.145	0.137
$\mathbf{e} = [x_{f2}, qm_{f1}]$	0.794	0.204	0.198	0.211
$\mathbf{e} = [x_{f2}, qm_{f2}]$	0.811	0.188	0.185	0.191
$\mathbf{e} = [x_{f2}, qm_{f1}, qm_{f2}]$	0.807	0.194	0.195	0.192
$\mathbf{e} = [x_{f1}, x_{f2}]$	*0.886*	*0.112*	*0.105*	*0.120*
$\mathbf{e} = [x_{f1}, x_{f2}, qm_{f1}]$	0.898	0.103	0.105	0.101
$\mathbf{e} = [x_{f1}, x_{f2}, qm_{f2}]$	0.899	0.101	0.100	0.101
$\mathbf{e} = [x_{f1}, x_{f2}, qm_{f1}, qm_{f2}]$	0.890	0.110	0.110	0.110
$SVM - rbf$				
$\mathbf{e} = [x_{f1}, qm_{f1}]$	0.879	0.122	0.125	0.118
$\mathbf{e} = [x_{f1}, qm_{f2}]$	0.869	0.130	0.130	0.131
$\mathbf{e} = [x_{f1}, qm_{f1}, qm_{f2}]$	0.893	0.108	0.110	0.105
$\mathbf{e} = [x_{f2}, qm_{f1}]$	0.804	0.195	0.190	0.199
$\mathbf{e} = [x_{f2}, qm_{f2}]$	0.807	0.192	0.190	0.195
$\mathbf{e} = [x_{f2}, qm_{f1}, qm_{f2}]$	0.846	0.154	0.153	0.155
$\mathbf{e} = [x_{f1}, x_{f2}]$	*0.889*	*0.111*	*0.110*	*0.112*
$\mathbf{e} = [x_{f1}, x_{f2}, qm_{f1}]$	0.911	0.089	0.090	0.088
$\mathbf{e} = [x_{f1}, x_{f2}, qm_{f2}]$	0.907	0.094	0.095	0.093
$\mathbf{e} = [x_{f1}, x_{f2}, qm_{f1}, qm_{f2}]$	0.932	0.067	0.065	0.070

Table 5.6: $Q - stack$ results for face modality, development data set. The results obtained for the baseline systems are marked in italic font.

	AC	$HTER$	ER_A	ER_B
baseline				
$\mathbf{e} = x_{f1}$	*0.862*	*0.144*	*0.152*	*0.135*
$\mathbf{e} = x_{f2}$	*0.690*	*0.270*	*0.209*	*0.331*
$SVM - lin$				
$\mathbf{e} = [x_{f1}, qm_{f1}]$	**0.872**	**0.136**	**0.149**	**0.123**
$\mathbf{e} = [x_{f1}, qm_{f2}]$	0.854	0.145	0.143	0.147
$\mathbf{e} = [x_{f1}, qm_{f1}, qm_{f2}]$	**0.860**	**0.135**	**0.127**	**0.143**
$\mathbf{e} = [x_{f2}, qm_{f1}]$	**0.749**	**0.244**	**0.233**	**0.255**
$\mathbf{e} = [x_{f2}, qm_{f2}]$	**0.799**	**0.214**	**0.233**	**0.195**
$\mathbf{e} = [x_{f2}, qm_{f1}, qm_{f2}]$	**0.768**	**0.217**	**0.195**	**0.240**
$\mathbf{e} = [x_{f1}, x_{f2}]$	*0.859*	*0.131*	*0.115*	*0.147*
$\mathbf{e} = [x_{f1}, x_{f2}, qm_{f1}]$	**0.880**	**0.126**	**0.137**	**0.116**
$\mathbf{e} = [x_{f1}, x_{f2}, qm_{f2}]$	**0.870**	**0.122**	**0.108**	**0.135**
$\mathbf{e} = [x_{f1}, x_{f2}, qm_{f1}, qm_{f2}]$	**0.890**	**0.115**	**0.122**	**0.107**
$Bayes$				
$\mathbf{e} = [x_{f1}, qm_{f1}]$	**0.877**	**0.140**	**0.166**	**0.114**
$\mathbf{e} = [x_{f1}, qm_{f2}]$	**0.858**	**0.141**	**0.138**	**0.143**
$\mathbf{e} = [x_{f1}, qm_{f1}, qm_{f2}]$	**0.884**	**0.132**	**0.155**	**0.108**
$\mathbf{e} = [x_{f2}, qm_{f1}]$	**0.670**	**0.265**	**0.165**	**0.365**
$\mathbf{e} = [x_{f2}, qm_{f2}]$	**0.768**	**0.212**	**0.180**	**0.243**
$\mathbf{e} = [x_{f2}, qm_{f1}, qm_{f2}]$	**0.753**	**0.219**	**0.177**	**0.262**
$\mathbf{e} = [x_{f1}, x_{f2}]$	*0.867*	*0.128*	*0.120*	*0.136*
$\mathbf{e} = [x_{f1}, x_{f2}, qm_{f1}]$	0.875	0.129	0.136	0.122
$\mathbf{e} = [x_{f1}, x_{f2}, qm_{f2}]$	**0.875**	**0.117**	**0.105**	**0.129**
$\mathbf{e} = [x_{f1}, x_{f2}, qm_{f1}, qm_{f2}]$	**0.884**	**0.117**	**0.118**	**0.116**
$SVM - rbf$				
$\mathbf{e} = [x_{f1}, qm_{f1}]$	0.844	0.155	0.153	0.157
$\mathbf{e} = [x_{f1}, qm_{f2}]$	0.857	0.144	0.145	0.143
$\mathbf{e} = [x_{f1}, qm_{f1}, qm_{f2}]$	0.847	0.151	0.148	0.154
$\mathbf{e} = [x_{f2}, qm_{f1}]$	**0.748**	**0.254**	**0.257**	**0.251**
$\mathbf{e} = [x_{f2}, qm_{f2}]$	**0.785**	**0.217**	**0.221**	**0.214**
$\mathbf{e} = [x_{f2}, qm_{f1}, qm_{f2}]$	**0.771**	**0.230**	**0.231**	**0.229**
$\mathbf{e} = [x_{f1}, x_{f2}]$	*0.865*	*0.134*	*0.131*	*0.136*
$\mathbf{e} = [x_{f1}, x_{f2}, qm_{f1}]$	0.851	0.148	0.147	0.150
$\mathbf{e} = [x_{f1}, x_{f2}, qm_{f2}]$	**0.878**	**0.123**	**0.123**	**0.122**
$\mathbf{e} = [x_{f1}, x_{f2}, qm_{f1}, qm_{f2}]$	**0.870**	**0.132**	**0.135**	**0.129**

Table 5.7: $Q - stack$ results for face modality, evaluation data set. The results obtained for the baseline systems are marked in italic font. The results for classification with quality measures in the proposed $Q - stack$ scheme which are superior to those obtained using baseline systems are marked in bold font.

Discussion of the results

The baseline classifiers for the face modality had a difficult task - one-to-one face image matching based only on one single training image is a challenging enterprise [106]. Most known experimental protocols designed for the face verification task assume having multiple images for the user model training [10, 13]. This is of particular importance for the subspace projection methods, including the PCA classifier used in the experiments reported here, which are known to be sensitive to geometrical and photometrical variations in classified face images [106, 168]. Hence the clearly weaker performance of the baseline PCA baseline system in comparison with the baseline DCT system. It must be stressed here that the error rates obtained using a difficult experimental protocol defined on a database containing low-quality face images are realistic, in the sense that they are comparable to the error rates reported from studies in real, out-of-the-lab environments [114]. However, we reiterate here that rather than optimizing the performance of the base classifiers the main goal of the work was to show the improvements in classification due to the use of quality-independent information in the $Q - stack$ framework.

An analysis of the dependence estimates between scores and quality measures for the DCT classifier reveals that the relationships between x_{f1} and qm_{f1} are more consistent across classes than those between x_{f1} and qm_{f2}. Namely, the correlation coefficient between x_{f1} and qm_{f2} changes sign between classes A and B. Naturally this fact is not reflected in the mutual information estimates. The difference between the correlation coefficients between x_{f1} and qm_{f1} for the development and evaluation data sets is also significant, which suggests that the data samples were not sufficiently representative of the population. Consequently, the dependence between x_{f1} and qm_{f1} appears more pronounced than that between x_{f1} and qm_{f2}, although the average estimates for both classes do not promise particularly high gains in classification accuracy. Out of the two, qm_{f1} which sports more consistent dependence on x_{f1} than qm_{f2} does, appears to be a better choice for $Q - stack$. These observations are is indeed confirmed by the experimental results listed in Table 5.7.

For the PCA classifier the strong and consistent dependence between x_{f2} and qm_{f2} suggests tangible improvements in classification performance for $Q - stack$ in comparison with the baseline. This strong dependence is also clearly visible in figures 5.15, 5.13 and 5.14. The corresponding error rates listed in Table 5.7 are significantly lower for $Q - stack$ than for the baseline classifier.

At the same time, however, both qm_{f1} and qm_{f2} show strong and consistent dependence on one another, reflected in high-valued and consistent mutual information and correlation estimates, which hints on potential benefits out of using them both with either x_{f1} or x_{f2}, or both. This is confirmed by the experimental results, where $Q - stack$ applied to single- as well as multiple-classifier face matching delivered better performance than the baseline systems.

The presented figures and classification results shown in Table 5.7 clearly suggest that the choice of an appropriate stacked classifier is of the essence for harvesting maximal benefits from the particular configuration of evidence in $Q - stack$. The overall best performance was obtained using the linear $SVM - lin$ classifier, which is most restrictive terms of complexity of its decision hyperplane. This can be attributed to the fact that the database offered rather limited size of the development set which made estimation of free parameters for the remaining two stacked classifiers less reliable. It is particularly visible for instance in figures 5.8 and 5.11, where the decision boundary is clearly overtrained. On the other hand the complexity of the decision boundary created by the *Bayes* stacked classifier shown in Figure 5.15 is well suited for the structure of the classification problem, and the resulting classification accuracy is higher than that for the other used stacked classifiers.

The graphical representations of the class-conditional distributions of evidence shown in Figures 5.6 through 5.14 help to intuitively understand how the dependence between the scores and quality measures, quantified in Tables 5.4 and 5.5, impacts the class separation and consequently obtained

classification results for the face modality. Let us consider Figure 5.13, where classification in the evidence space defined by $\mathbf{e} = [x_{f2}, qm_{f2}]$ is shown. As the result of the strong positive correlation between x_{f2} and qm_{f2} the class-conditional joint distributions of $e|A$ and $e|B$ are visibly *slanting* to the right for higher values of qm_{f2}. As a result, the optimal class decision boundary differs significantly from that defined by the baseline score threshold, τ_{f2}, and the classes are being separated better by the stacked classifier than it is the case of the baseline classifier. Consequently, the classification error rates, shown in Table 5.7, are clearly much lower for $\mathbf{e} = [x_{f2}, qm_{f2}]$ than for the baseline of $\mathbf{e} = [x_{f2}]$. The reader is encouraged to follow the same analysis for other graphical representations of classification in the evidence space shown in respective figures in this chapter.

The experiments for the face modality show that for all considered baseline systems adding quality measures as additional components of the evidence vector in the proposed method of $Q-stack$ proved to be beneficial from the viewpoint of classification performance, in single- and multi classifier scenarios.

5.2.4 Error prediction and rejection

In this section we present the results of error prediction technique discussed in Chapter 4, applied to face verification. We consider these two single-classifier scenarios where one of the PCA and DCT classifiers are used, and the multiple-classifier scenario where both classifiers are used together. Using procedures described in Chapter 4 credence estimates are computed for each evidence combination, and then applied to rejecting potentially unreliable classification decisions whose credence falls below a set credence threshold. The error prediction performance is evaluated using the monotonicity and accountability criteria discussed in detail in Section 4.2.3.

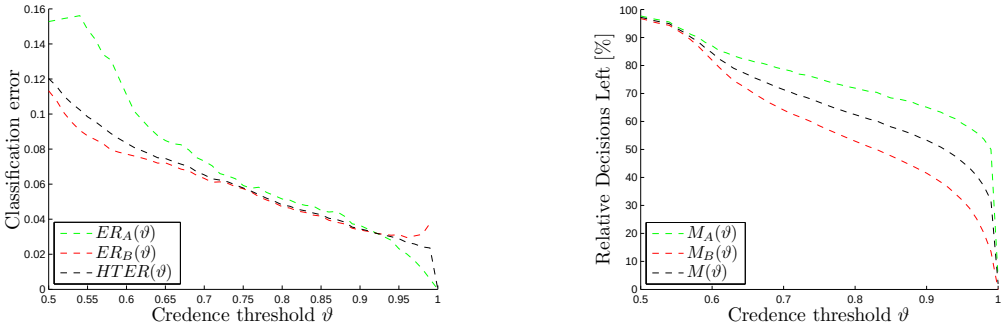
For the face modality, following evidence combinations were used in the error prediction experiments:

- Single classifier, DCT: $\mathbf{e} = [x_{f1}]$, $\mathbf{e} = [x_{f1}, qm_{f1}]$.
- Single classifier, PCA: $\mathbf{e} = [x_{f2}]$, $\mathbf{e} = [x_{f2}, qm_{f2}]$.
- Multiclassifier fusion, $DCT + PCA$: $\mathbf{e} = [x_{f1}, x_{f2}]$, $\mathbf{e} = [x_{f1}, x_{f2}, qm_{f1}, qm_{f2}]$.

For single-classifier architectures, Figures 5.19, 5.20, 5.21 and 5.22 show the accuracy gains from credence thresholding, subplots (a), and relative number of classification decision remaining after credence thresholding, subplots (b). Results obtained with the use of relevant quality measures are shown in Figures 5.20 and 5.22, and those obtained using scores alone are presented in Figures 5.19 and 5.21. The shape of the curves in subplots (a) allows the reader to evaluate the credence estimates in terms of the monotonicity criterion. A good credence estimator should result in a monotonically growing total accuracy, as discussed in Section 4.2.1.

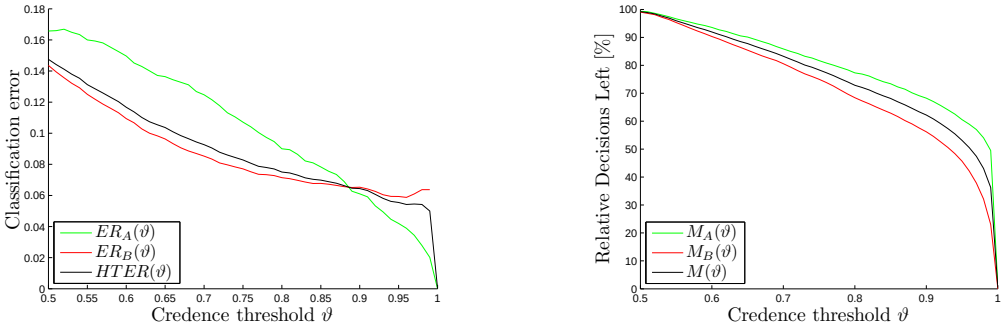
The results of error prediction are represented in a similar fashion for multi-classifier face matching in figures 5.23 (without the use of quality measures, $\mathbf{e} = [x_{f1}, x_{f2}]$) and 5.24 (with the use of quality measures, $\mathbf{e} = [x_{f1}, x_{f2}, x_{qm1}, x_{qm2}]$).

Numerical results gathered in Table 5.8 allow for an evaluation of the credence estimates according to the accountability criterion. For each considered configuration of evidence, the mean observed classification accuracy AC, $HTER$, and per-class errors ER_A and ER_B are reported together with their standard deviations after 100 experimental runs. Corresponding mean credence estimators with their standard deviations are given. The differences between the mean observed values and the mean credence predictors are reported (*Mean difference* δ). According to the accountability criterion, the smaller the value of δ the more accurate the error prediction, and consequently the better the credence estimator.



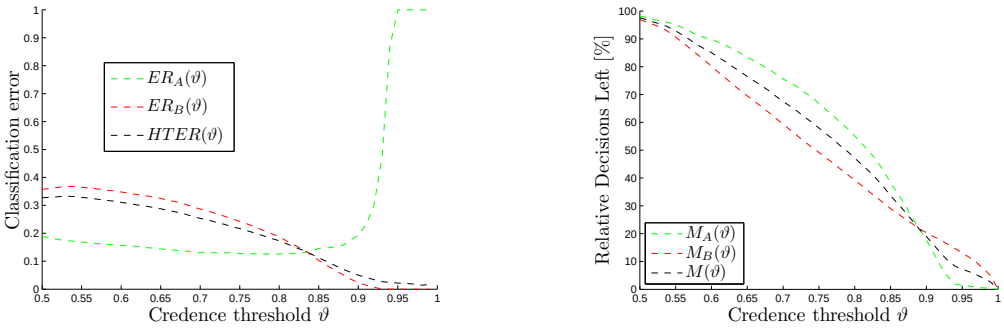
(a) error after credence thresholding (b) percentage of decisions after credence thresholding

Figure 5.19: Error prediction and rejection using credence estimates, face modality, *DCT* classifier, evidence configuration $\mathbf{e} = [x_{f1}]$.



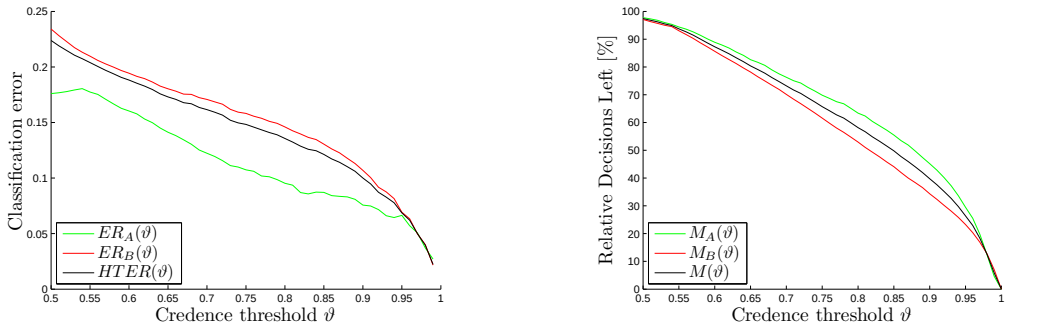
(a) error after credence thresholding (b) percentage of decisions after credence thresholding

Figure 5.20: Error prediction and rejection using credence estimates, face modality, *DCT* classifier, evidence configuration $\mathbf{e} = [x_{f1}, qm_{f1}]$.



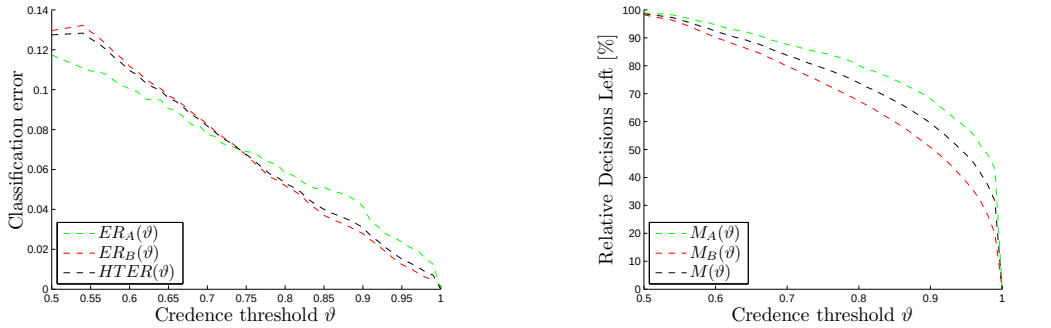
(a) error after credence thresholding (b) percentage of decisions after credence thresholding

Figure 5.21: Error prediction and rejection using credence estimates, face modality, *PCA* classifier, evidence configuration $\mathbf{e} = [x_{f2}]$.



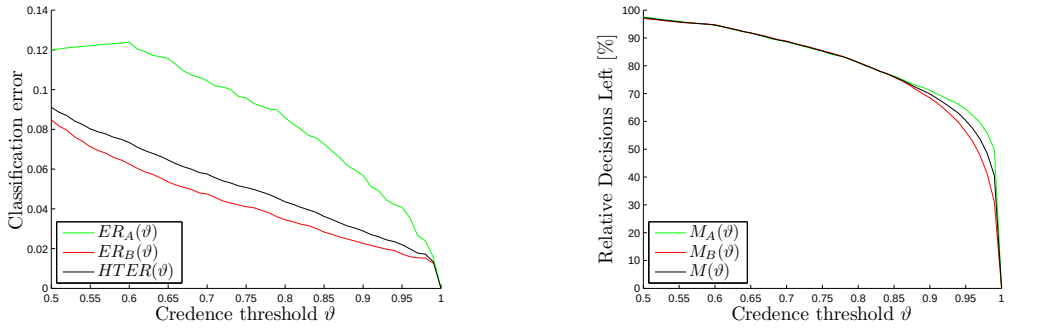
(a) error after credence thresholding (b) percentage of decisions after credence thresholding

Figure 5.22: Error prediction and rejection using credence estimates, face modality, PCA classifier, evidence configuration $\mathbf{e} = [x_{f2}, qm_{f2}]$.



(a) error after credence thresholding (b) percentage of decisions after credence thresholding

Figure 5.23: Error prediction and rejection using credence estimates, face modality, multi-classifier fusion of $DCT + PCA$ classifiers, evidence configuration $\mathbf{e} = [x_{f1}, x_{f2}]$.



(a) error after credence thresholding (b) percentage of decisions after credence thresholding

Figure 5.24: Error prediction and rejection using credence estimates, face modality, multi-classifier fusion of $DCT + PCA$ classifiers, evidence configuration $\mathbf{e} = [x_{f1}, x_{f2}, qm_{f1}, qm_{f2}]$.

	AC, $\overline{R}$	$HTER$, $\overline{R_w}$	ER_A, $1-\overline{R_B}$	ER_B, $1-\hat{R_B}$
$\mathbf{e} = [x_{f1}]$				
Actual accuracy	0.857	0.146	0.149	0.142
St. dev. of actual accuracy	0.000	0.000	0.000	0.000
Credence estimate	0.822	0.153	0.192	0.113
St. dev. of credence estimate	0.000	0.000	0.000	0.000
Mean difference δ	**0.035**	**-0.007**	**-0.043**	**0.029**
$\mathbf{e} = [x_{f1}, qm_{f1}]$				
Actual accuracy	0.879	0.132	0.149	0.115
St. dev. of actual accuracy	0.002	0.001	0.003	0.003
Credence estimate	0.860	0.130	0.145	0.116
St. dev. of credence estimate	0.001	0.000	0.002	0.001
Mean difference δ	**0.018**	**0.002**	**0.005**	**-0.001**
$\mathbf{e} = [x_{f2}]$				
Actual accuracy	0.688	0.272	0.209	0.334
St. dev. of actual accuracy	0.010	0.003	0.008	0.014
Credence estimate	0.724	0.253	0.288	0.217
St. dev. of credence estimate	0.002	0.001	0.002	0.001
Mean difference δ	**-0.037**	**0.019**	**-0.079**	**0.117**
$\mathbf{e} = [x_{f2}, qm_{f2}]$				
Actual accuracy	0.752	0.220	0.176	0.264
St. dev. of actual accuracy	0.004	0.002	0.002	0.005
Credence estimate	0.811	0.184	0.192	0.176
St. dev. of credence estimate	0.002	0.005	0.001	0.009
Mean difference δ	**-0.059**	**0.036**	**-0.016**	**0.088**
$\mathbf{e} = [x_{f1}, x_{f2}]$				
Actual accuracy	0.873	0.127	0.127	0.127
St. dev. of actual accuracy	0.000	0.000	0.000	0.000
Credence estimate	0.859	0.124	0.150	0.097
St. dev. of credence estimate	0.000	0.000	0.000	0.000
Mean difference δ	**0.014**	**0.003**	**-0.024**	**0.031**
$\mathbf{e} = [x_{f1}, x_{f2}, qm_{f1}, qm_{f2}]$				
Actual accuracy	0.884	0.117	0.118	0.116
St. dev. of actual accuracy	0.001	0.001	0.002	0.001
Credence estimate	0.885	0.107	0.119	0.094
St. dev. of credence estimate	0.002	0.001	0.002	0.001
Mean difference δ	**-0.001**	**0.010**	**-0.001**	**0.022**

Table 5.8: Evaluation of credence estimates using the accountability criterion. The mean difference between actual observed error and the mean credence estimates after 100 experimental iterations is given by δ

Discussion of the results

Figures 5.19, 5.21, 5.20, 5.22, 5.23 and 5.24 show that classification error expressed in terms of $HTER$, ER_A and ER_B decreases with the growing value of the credence threshold ϑ. This behavior is in accord with the theoretical expectations discussed in Section 4.2.1, and it agrees with the intuitive notion of the concept of decision credence: discarding classification decisions deemed not reliable in terms of its credence estimate indeed resulted in reduced observed error over the remaining decisions. This effect proves that the relative number of erroneous decisions in the discarded pool of decisions was higher than that in the decisions considered credible according to the credence estimates. In all presented graphs $HTER$ decreases monotonically as ϑ grows, which satisfies the monotonicity criterion for evaluation of the credence estimates.

Note that in Figure 5.21 the curve representing the change of ER_A as a function of ϑ is not monotonic, and neither is the shape of ER_B in Figure 5.20, although here the effect is far less pronounced. In both of these cases the class error rates for one class decrease with growing credence threshold until they reach certain value of ϑ from which on they begin to grow. This effect is due to the imperfectness of distribution modeling and has been discussed in section 4.2.1. Namely, if the maximal credence estimates for one of the classes is $R_{max} < 1$ then most of the potentially erroneous decisions will have been discarded before the actual credence threshold reaches $\vartheta = R_{max}$. Due to the fact that $R_{max} \ll 1$ most decisions for this class that are still remaining with credence $R > R_{max}$ are actual misclassifications, hence growing class error as ϑ grows from $\vartheta = R_{max}$ to $\vartheta = 1$. However, this effect concerns only a relatively small number of misclassified decisions, as it is visible from the curves showing the percentage of remaining decisions after credence thresholding. Therefore the monotonic declining tendency of $HTER(\vartheta)$ remains unaltered.

The second criterion in evaluation of the credence estimates is the accountability criterion, according to which the difference between the weighted sum of all credence estimates and the observed class-accuracy and $HTER$ must be minimized. These differences are listed in Table 5.8. The δ values reach maximal value of -5.9% for the evidence configuration of $\mathbf{e} = [x_{f2}, qm_{f2}]$, and the minimal value of -0.1% for the evidence combination of $\mathbf{e} = [x_{f1}, x_{f2}, qm_{f1}, qm_{f2}]$. Note that the high value of δ in the first case is due to a large discrepancy between the observed error and the mean credence estimates for class B, and that this effect coincides with the non-monotonic behavior of $ER_B(\vartheta)$ in Figure 5.21 discussed before. The explanation of this coincidence is that both effects result from the fact that for the particular models obtained for this evidence combination the credence estimates fail to reach a value close to 1.

In Chapter 4 we have applied the Kolmogorov-Smirnov test (KS) to show that the means of credence estimates and the observed classification accuracies are samples of the same distribution, for the synthetic data example. In the practical example considered in this chapter using the KS test is not meaningful, for, as the Figure 5.3 shows, the actual observed evidence distributions of the development and evaluation data sets differ visibly. Therefore we consider the actual value of δ as a measure of error in credence estimation. Since in all considered cases the value of δ is much smaller than the actual error rates, we conclude that the proposed credence estimation method returned satisfyingly accurate results.

The presented results show that the presented method of credence estimation allows for accurate error prediction and rejection in face verification in the single- and multiple-classifier scenarios, in both cases with- and without an involvement of quality measures.

5.3 Fingerprint Modality

For the fingerprint modality, two separate acquisition channels were used: an optical sensor and a thermal sensor. For data acquisition via each of the channels following devices were used.

- Thermal scanner: Atmel AT77C104B fingerprint sensor dedicated to PDA, cellular and smartphone applications. Based on FingerChip® thermal technology, the AT77C104B is a linear sensor that captures fingerprint images by sweeping the finger over the sensing area [1].
- Optical scanner: Biometrika FX 2000, stand-alone optical fingerprint scanner [3].

Examples of fingerprints collected using each of these devices are shown in Figure 5.1. In the experiments reported in this chapter the matching scores obtained for images collected using the optical sensor scanner are denoted as x_{p1} and the matching scores obtained for images collected using the thermal sensor scanner are denoted as x_{p2}*.

5.3.1 Baseline classifier

Fingerprint images acquired using both sensors have been matched using the publicly available automated fingerprint matching package NFIS2 [162]. Currently, the NFIS2 package is a standard benchmarking fingerprint matching software, used for instance in the evaluations of the U.S. Visit program [65, 163]. In the NFIS2 software package, fingerprint matching is based on the minutiae information. The minutiae extraction and consequent matching is performed by two dedicated components of the package, MINDTCT and BOZORTH3.

Minutiae extraction

Minutiae extraction (localization) in an input fingerprint is realized by the MINDTCT function. The function follows a sequence of routines, listed below [162]:

1. Input Fingerprint File: the function accepts ANSI/NIST, WSQ, JPEGB, JPEGL, and IHEAD file formats.
2. Generate Image Maps: determines the directional flow of ridges in the image and detects image regions of low contrast, low ridge flow, and high ridge curvature. Minutiae may be extracted unreliably from these regions.
3. Binarize Image: The image is binarized based on the ridge flow direction in the immediate neighborhood of the pixel. If in the presence of weak ridge flow in given pixel's neighborhood the pixel's binarized value is 1, otherwise it is set to 0.
4. Detect Minutiae: minutiae are detected in the binary image by comparing local subimages to predefined binary patterns, different for the ridge endings and for the ridge bifurcations.
5. Remove False Minutiae: performs heuristic removal of spurs, hooks, lakes, holes and other artifacts mistaken for minutiae by the detection routine.
6. Count Neighbor Ridges: counts the number of ridges intervening between the considered minutiae and its immediate neighbors (up to five neighbors).

*in the experiments reported here, fingerprint scores and quality measures are courtesy of Prof. Javier Garcia-Ortega and Dr. Julian Fierrez-Aguilar, Universidad Autonoma Madrid.

7. Assess Minutiae Quality: provides an individual estimate of quality of each detected minutiae based on its location in the original image and on the first and second order statistics of the pixels in the original image, in an 11-pixel-radius neighborhood.

8. Output Minutiae: the information concerning the extracted minutiae is saved to a file which can be further used by the BOZORTH3 fingerprint matcher.

The MINDTCT procedures have been designed and optimized to work best with 500dpi resolution, 8-bit grayscale fingerprint images, which makes them compatible with the images from the Biosec database.

Fingerprint matching

Prior to matching, the minutiae information about both matched fingerprints is extracted and saved in corresponding files by the MINDTCT routine. Thus created files are input to the BOZORTH3 matching routine which compares the minutiae information about both compared fingerprints contained in the input files, and returns a matching similarity score. The minutiae information used by the matcher is limited to location (x, y) and orientation t, represented as x, y, t. The algorithm is designed to be rotation and translation invariant.

The algorithm is comprised of three major steps:

1. Construct Intra-Fingerprint Minutia Comparison Tables: two tables containing all minutiae information for both matched fingerprints are created, as well as a minutiae compatibility table.

2. Construct an Inter-Fingerprint Compatibility Table: compare both fingerprint's minutia comparison tables and construct a compatibility table containing the minutiae pairs that are common for both compared fingerprints.

3. Traverse the Inter-Fingerprint Compatibility Table: compatible minutiae pairs are linked to form a graph. The longest path to traverse the graph is the theoretically simplest way of computing the matching score between two fingerprints, once the graph is constructed. In practice, several factors decided that a multiple-start routine for traversing the graph is used instead of a full graph search.

Detailed description of the NFIS2 system is beyond the scope and frame of this work. For exhaustive algorithm and implementation details the reader is referred to [162]. In the experiments reported in this thesis we used the NFIS2 for matching of images originating from both optical and thermal sensor. The matching scores obtained for the images originating from the optical scanner are denoted as x_{p1} and those originating from the thermal senor as x_{p2}.

5.3.2 Quality measures

In the experiments reported here we use two fingerprint quality measures: the NFIQ quality measures which is an integral part of the NFIS2 fingerprint matching software package [162], and a quality measure based on the spectral analysis of the fingerprint image, proposed in [29]. Both quality measures have been introduced in Chapter 2, Section 2.6.2, without quoting the algorithm details. Since understanding of the workings of the quality measurement is essential for the appreciation of the link between the quality estimates and the matching similarity scores, here we provide necessary details of both algorithms used in the experiments reported in this thesis.

Quality index in the frequency domain

The choice of this quality measure for our experiments was dictated by the fact that the authors of the proposed algorithm have shown that the resulting quality estimates predict well the success in detecting minutiae [29], and the fingerprint matcher used in our experiments is based on minutiae information. Therefore we anticipated that the used quality measure shall sport a dependence on the matching scores.

The fingerprint quality index computed in the frequency domain, proposed in [29], requires that the power spectrum P of a fingerprint image f of size $M \times N$ pixels is approximated:

$$P(k,l) \equiv |F(k,l)|^2 = |\frac{1}{NM}\sum_{a=1}^{N}\sum_{b=1}^{M} f(a,b)e^{-i2\pi(\frac{ka}{N}+\frac{lb}{M})}|^2, \tag{5.12}$$

where $i = \sqrt{-1}$, and $f(a,b)$ refers to the gray level pixel intensity at position (a,b) in the image f. Then, a set of bandpass filters, equally spaced in the frequency domain, is built:

$$R_t(k,l) = H(k,l|m_t+1,n) - H(k,l|m_t,n), \tag{5.13}$$

where $m_t = 0.06 + t\frac{0.5-0.06}{T}$, $t = 1,2,...,15$, and

$$H(k,l|m,n) = \frac{1}{1+\frac{1}{m^{2n}}((\frac{k-k_0}{M})^2+(\frac{l-l_0}{N})^2)^n} \tag{5.14}$$

is a Butterworth function which defines a low-pass filter with the cutoff frequency defined by m and filter order given by n. The values for parameters m_t and t are selected according to the observed spatial frequency structure of high quality fingerprints. In Equation 5.14 (k_0, l_0) is the location of the center of the power spectrum corresponding to spatial frequency $(0,0)$. From here, the energy concentrated in the t-th power spectrum band is computed by

$$E_t = \sum_{k=0}^{N-1}\sum_{l=0}^{M-1} R_t(k,l)P(k,l). \tag{5.15}$$

A high quality fingerprint is expected to have the energy concentrated in a narrow energy band, and therefore the corresponding distribution of E_t should have a well-pronounced peak. Consequently, the quality measure is computed as a difference between the maximal entropy $\hat{E}$ of a uniform discrete distribution with t bins and the actual entropy E of the distribution of energy over T frequency bands:

$$qm = \hat{E} - E = \hat{E} + \sum_{t=1}^{T} P_t log(P_t). \tag{5.16}$$

Here, $P_t = \frac{E_t}{\sum_T E_t}$ is normalized energy in frequency band t. Examples of fingerprints of high, medium, and low quality with their corresponding spatial power spectra and distributions of the bandpass Butterworth filter responses are shown in figure 5.25*.

When comparing two fingerprint images, their combined quality index is calculated as the geometrical means of quality measures computed for both images, as given by Equation 5.5. In the experimental results reported in this thesis, thus calculated fingerprint quality measures are denoted as qm_{p11} (computed for images obtained from the optical sensor) and qm_{p12} (computed for images obtained from the thermal sensor).

*Images are courtesy of Yi Chen, Michigan State University.

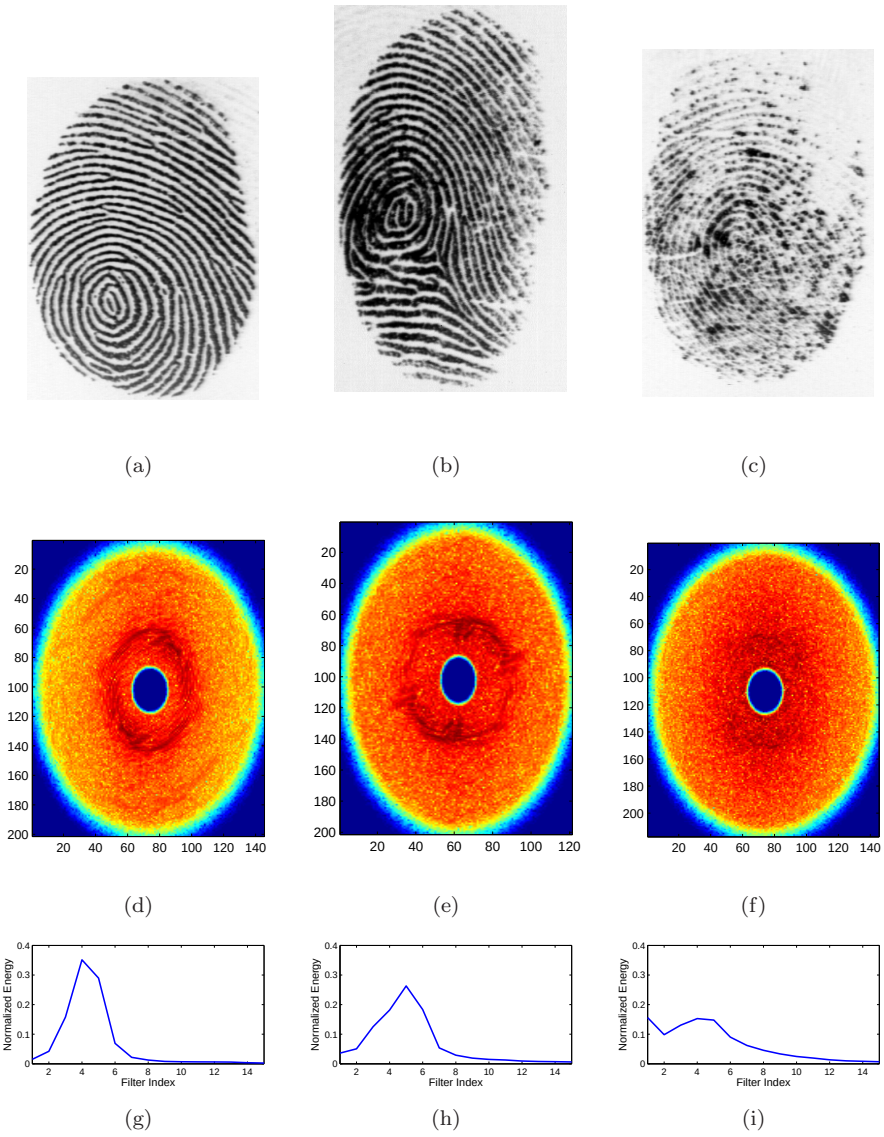


Figure 5.25: Measuring fingerprint quality in the spatial frequency domain: (a,b,c) - high-, medium-, and low-quality fingerprint examples, (d,e,f) - corresponding spatial power spectra, (g,h,i) - corresponding histograms of Butterworth filter responses. Images taken from [29].

NFIQ - Classification-based quality estimate

Adoption of the $NFIQ$ quality measure, which is a part of the $NFIS2$ package, is a consequence of the choice to use the $NFIS2$ fingerprint matching software in our experiments. The $NFIQ$ quality measure is especially designed to predict the variation in the obtained classification scores due to the fingerprint image quality. The $NFIQ$ quality measure, which we discussed in Chapter 2, Section 2.6.2, produces a discrete fingerprint quality estimate, with five possible discrete levels. The assignment of a quality label to a considered fingerprint sample is done via a classification step, in which image quality features are processed by a pre-trained neural network classifier. Quality of level 1 corresponds to best, and level 5 to the worst fingerprint quality with respect to the development set used in the training of the neural network quality classifier. The details of implementation of the quality measurement algorithm are beyond the frames of this thesis, but can be readily accessed from [150, 162].

Similarly as with the other quality measures, the combined quality index for a pair of compared fingerprints is calculated as the geometrical mean of quality measures computed for both samples (Equation 5.5). As a consequence of this, the resulting combined quality measure has 10 discrete levels. In this chapter, the combined $NFIQ$ quality measures are denoted as qm_{p21} for images obtained from the optical sensor and qm_{p22} for images obtained from the thermal sensor.

5.3.3 $Q-stack$ in fingerprint verification

The experimental results for fingerprint modality are reported according to the order spelled out in Section 5.1.3. The available evidence components for the fingerprint modality include:

Optical sensor Similarity scores x_{p1}, quality measures qm_{p11} and qm_{p12}.

Thermal sensor Similarity scores x_{p2}, quality measures qm_{p21} and qm_{p22}.

In the following section we provide a comparison of baseline classification results obtained for the fingerprint modality with the results obtained by using the proposed method of $Q-stack$. The baseline results are obtained for:

Single classifier systems Two single-classifier systems are considered that use matching scores x_{p1} and x_{p2}, respectively.

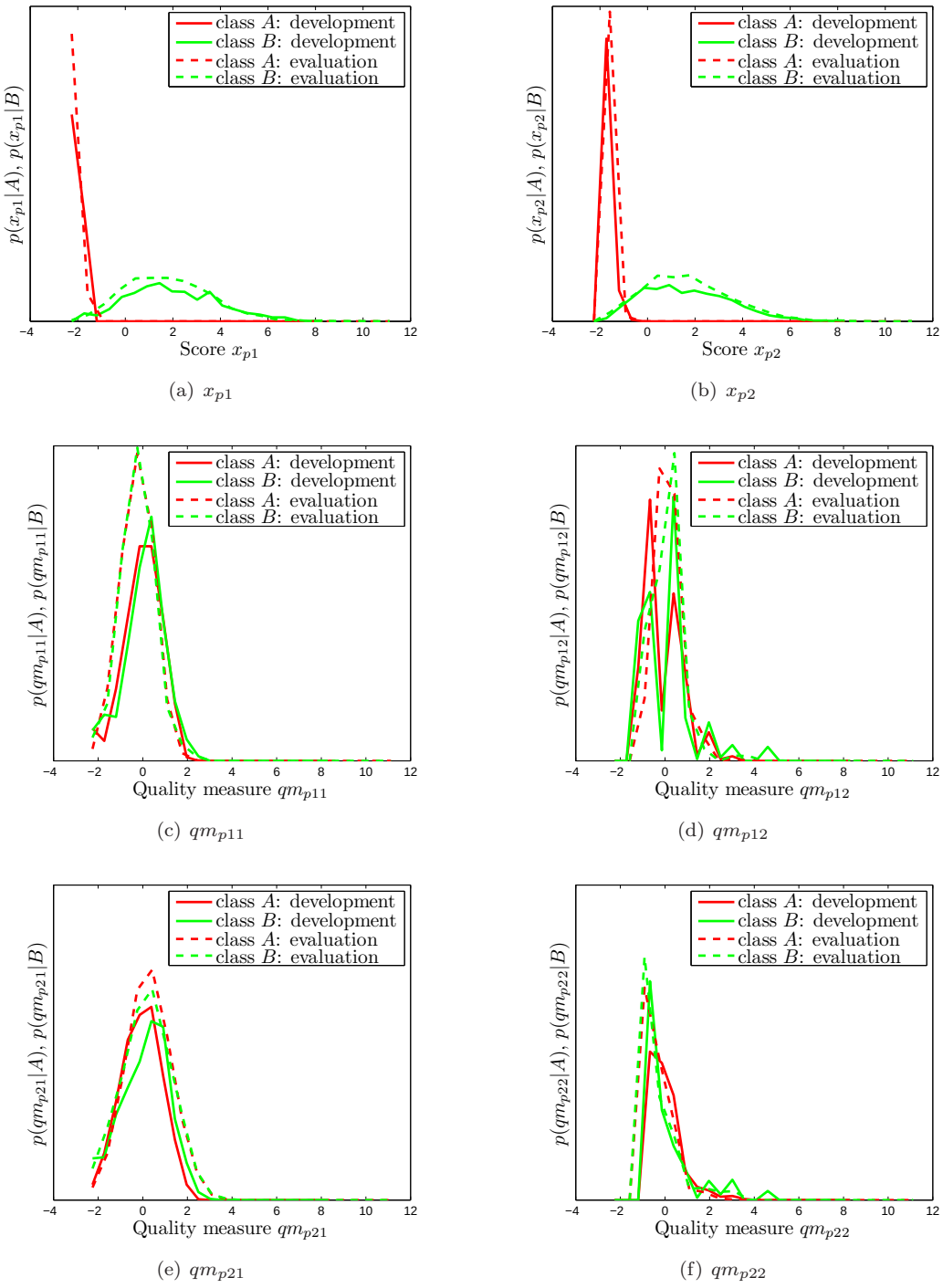
Multimodal system A multimodal system is considered that uses the scores originating from both fingerprint modalities, x_{f1} and x_{f2}.

We present the experimental results following the steps listed in Section 5.1.3.

Analysis of dependencies between evidence components

The marginal class-conditional distributions for all available evidence component are shown in Figure 5.26. As a consequence of the fact that qm_{p12} and qm_{p22} take discrete values, their resulting class-conditional marginal distributions are not smooth like those of qm_{p11} and qm_{p21}. The shape of the obtained class-conditional score distributions follow closely the distributions resulting from applying the same $NFIS2$ fingerprint matcher to other databases, as reported in [149, 150], which hints on the fact that the results reported here are probably also valid for other publicly available fingerprint databases.

The complicated shape of the marginal distributions of quality measures makes it difficult to visually evaluate to what extent are the quality measures indeed class-independent. For this reason



(a) x_{p1} (b) x_{p2}

(c) qm_{p11} (d) qm_{p12}

(e) qm_{p21} (f) qm_{p22}

Figure 5.26: Class-conditional evidence distributions $p(e|A)$ and $p(e|B)$, fingerprint modality.

in Table 5.9 we provide the classification results for the evaluation data set obtained using exclusively the quality measures, from where it is evident that indeed the quality measures alone are individually irrelevant classification features and allow only for classification at an error rate close to the chance level of $HTER = 0.5$.

	AC	$HTER$	ER_A	ER_B
development				
$\mathbf{e} = [qm_{p11}]$	0.527	0.473	0.475	0.471
$\mathbf{e} = [qm_{p12}]$	0.564	0.439	0.456	0.422
$\mathbf{e} = [qm_{p21}]$	0.549	0.453	0.461	0.444
$\mathbf{e} = [qm_{p22}]$	0.402	0.578	0.486	0.670
evaluation				
$\mathbf{e} = [qm_{p11}]$	0.616	0.502	0.685	0.319
$\mathbf{e} = [qm_{p12}]$	0.513	0.476	0.459	0.493
$\mathbf{e} = [qm_{p21}]$	0.456	0.516	0.472	0.559
$\mathbf{e} = [qm_{p22}]$	0.470	0.534	0.540	0.528

Table 5.9: Classification of the samples from the evaluation data set with quality measures alone, fingerprint modality.

In order to gain a good comprehension of the dependence relationships between all the evidence components, Pearson's correlation coefficients (Equation 5.1) and mutual information (Equation 5.2) estimates computed pair-wise for all scores and quality measure of the fingerprint modality are given in Tables 5.10, 5.12 (development set), and 5.11, 5.13 (evaluation set). The results are reported for each class individually (A,B), for the entire data set, ($A+B$), and as an average of the dependence estimates for each class ($\overline{A, B}$).

As the results in Tables 5.10, 5.12, 5.11 and 5.13 show, for the entire data set ($A + B$) the estimator based on mutual information hints on stronger dependencies in the data than it is the case of the correlation-based estimator. Since the evidence distributions, particularly those of the quality measures as shown in Figure 5.26, are strongly non-gaussian, it can be assumed that the mutual information estimates reflect the actual data dependencies better than the linear correlation coefficients. Further, considering that a pronounced dependence exists between the matching scores and quality measures, gains in classification performance can be expected from augmenting the evidence space with the quality measure dimensions. Here, the frequency domain quality measure (Section 5.3.2) appears to offer a better potential of improving class separation based on its relatively high dependence estimates on both matching scores x_{p1} and x_{p2}.

Classification in the evidence space

In this section we present the classification results produced by three considered stacked classifiers in the proposed $Q-stack$ method, and compare them with the results obtained for the baseline systems with decision thresholds τ_{p1} and τ_{p2} for the optical and thermal modalities, respectively. Again, for the purpose of clarity the visualization of the $Q - stack$ classification boundaries is presented only where the evidence vector contained two components. In all figures the decision boundaries obtained by each particular stacked classifier are consistently shown in different colors, namely: $SVM - lin$ - black; $SVM - rbf$ - magenta; $Bayes$: blue. In each of the presented Figures corresponding baseline thresholds $x = \tau_{p1}$ and $x = \tau_{p2}$ are marked by dashed lines.

Figures 5.39, 5.40 and 5.41 show classification in the evidence space $\mathbf{e} = [x_{p1}, x_{p2}]$, which is

A						
	x_{p1}	qm_{p11}	qm_{p12}	x_{p2}	qm_{p21}	qm_{p22}
x_{p1}	1.00	0.12	0.05	0.13	0.13	-0.01
qm_{p11}	0.12	1.00	-0.52	-0.02	0.29	-0.54
qm_{p12}	0.05	-0.52	1.00	0.15	-0.16	0.45
x_{p2}	0.13	-0.02	0.15	1.00	0.17	-0.06
qm_{p21}	0.13	0.29	-0.16	0.17	1.00	-0.59
qm_{p22}	-0.01	-0.54	0.45	-0.06	-0.59	1.00
B						
	x_{p1}	qm_{p11}	qm_{p12}	x_{p2}	qm_{p21}	qm_{p22}
x_{p1}	1.00	0.66	-0.31	0.38	0.37	-0.44
qm_{p11}	0.66	1.00	-0.60	0.29	0.45	-0.59
qm_{p12}	-0.31	-0.60	1.00	-0.11	-0.22	0.61
x_{p2}	0.38	0.29	-0.11	1.00	0.63	-0.36
qm_{p21}	0.37	0.45	-0.22	0.63	1.00	-0.56
qm_{p22}	-0.44	-0.59	0.61	-0.36	-0.56	1.00
$A+B$						
	x_{p1}	qm_{p11}	qm_{p12}	x_{p2}	qm_{p21}	qm_{p22}
x_{p1}	1.00	0.26	-0.05	0.81	0.20	-0.15
qm_{p11}	0.26	1.00	-0.55	0.13	0.37	-0.56
qm_{p12}	-0.05	-0.55	1.00	0.02	-0.18	0.54
x_{p2}	0.81	0.13	0.02	1.00	0.32	-0.13
qm_{p21}	0.20	0.37	-0.18	0.32	1.00	-0.57
qm_{p22}	-0.15	-0.56	0.54	-0.13	-0.57	1.00
$\overline{(A,B)}$						
	x_1	qm_{p11}	qm_{p12}	x_2	qm_{p21}	qm_{p22}
x_1	1.00	0.39	-0.13	0.25	0.25	-0.23
qm_{p11}	0.39	1.00	-0.56	0.14	0.37	-0.56
qm_{p12}	-0.13	-0.56	1.00	0.02	-0.19	0.53
x_2	0.25	0.14	0.02	1.00	0.40	-0.21
qm_{p21}	0.25	0.37	-0.19	0.40	1.00	-0.58
qm_{p22}	-0.23	-0.56	0.53	-0.21	-0.58	1.00

Table 5.10: Correlation coefficients between the components of the evidence vector, fingerprint, development set.

A						
	x_{p1}	qm_{p11}	qm_{p12}	x_{p2}	qm_{p21}	qm_{p22}
x_{p1}	1.00	0.68	-0.36	0.38	0.35	-0.29
qm_{p11}	0.68	1.00	-0.46	0.34	0.41	-0.40
qm_{p12}	-0.36	-0.46	1.00	-0.19	-0.30	0.50
x_{p2}	0.38	0.34	-0.19	1.00	0.58	-0.34
qm_{p21}	0.35	0.41	-0.30	0.58	1.00	-0.44
qm_{p22}	-0.29	-0.40	0.50	-0.34	-0.44	1.00
B						
	x_{p1}	qm_{p11}	qm_{p12}	x_{p2}	qm_{p21}	qm_{p22}
x_{p1}	1.00	0.50	-0.20	0.26	0.21	-0.16
qm_{p11}	0.50	1.00	-0.38	0.17	0.25	-0.13
qm_{p12}	-0.20	-0.38	1.00	-0.18	-0.19	0.19
x_{p2}	0.26	0.17	-0.18	1.00	0.58	-0.32
qm_{p21}	0.21	0.25	-0.19	0.58	1.00	-0.39
qm_{p22}	-0.16	-0.13	0.19	-0.32	-0.39	1.00
$A+B$						
	x_{p1}	qm_{p11}	qm_{p12}	x_{p2}	qm_{p21}	qm_{p22}
x_{p1}	1.00	0.13	-0.08	0.84	0.03	-0.03
qm_{p11}	0.13	1.00	-0.32	0.07	0.23	-0.22
qm_{p12}	-0.08	-0.32	1.00	-0.06	-0.22	0.34
x_{p2}	0.84	0.07	-0.06	1.00	0.11	-0.06
qm_{p21}	0.03	0.23	-0.22	0.11	0.00	-0.32
qm_{p22}	-0.03	-0.22	0.34	-0.06	-0.32	1.00
$\overline{(A,B)}$						
	x_{p1}	qm_{p11}	qm_{p12}	x_{p2}	qm_{p21}	qm_{p22}
x_{p1}	1.00	0.59	-0.28	0.32	0.28	-0.22
qm_{p11}	0.59	1.00	-0.42	0.26	0.33	-0.26
qm_{p12}	-0.28	-0.42	1.00	-0.18	-0.24	0.35
x_{p2}	0.32	0.26	-0.18	1.00	0.58	-0.33
qm_{p21}	0.28	0.33	-0.24	0.58	1.00	-0.41
qm_{p22}	-0.22	-0.26	0.35	-0.33	-0.41	1.00

Table 5.11: Correlation coefficients between components of the evidence vector, fingerprint, evaluation set.

A						
	x_{p1}	qm_{p11}	qm_{p12}	x_{p2}	qm_{p21}	qm_{p22}
x_{p1}	-	0.00	0.01	0.01	0.00	0.01
qm_{p11}	0.00	-	0.09	0.00	0.10	0.11
qm_{p12}	0.01	0.09	-	0.01	0.10	0.06
x_{p2}	0.01	0.00	0.01	-	0.01	0.02
qm_{p21}	0.00	0.10	0.10	0.01	-	0.37
qm_{p22}	0.01	0.11	0.06	0.02	0.37	-
B						
	x_{p1}	qm_{p11}	qm_{p12}	x_{p2}	qm_{p21}	qm_{p22}
x_{p1}	-	0.50	0.14	0.28	0.33	0.23
qm_{p11}	0.50	-	0.19	0.27	0.31	0.29
qm_{p12}	0.14	0.19	-	0.15	0.27	0.23
x_{p2}	0.28	0.27	0.15	-	0.42	0.23
qm_{p21}	0.33	0.31	0.27	0.42	-	0.38
qm_{p22}	0.23	0.29	0.23	0.23	0.38	-
A+B						
	x_{p1}	qm_{p11}	qm_{p12}	x_{p2}	qm_{p21}	qm_{p22}
x_{p1}	-	0.19	0.06	0.93	0.13	0.12
qm_{p11}	0.19	-	0.09	0.09	0.13	0.14
qm_{p12}	0.06	0.09	-	0.06	0.14	0.10
x_{p2}	0.93	0.09	0.06	-	0.16	0.11
qm_{p21}	0.13	0.13	0.14	0.16	-	0.31
qm_{p22}	0.12	0.14	0.10	0.11	0.31	-
$\overline{(A,B)}$						
	x_{p1}	qm_{p11}	qm_{p12}	x_{p2}	qm_{p21}	qm_{p12}
x_{p1}	-	0.25	0.08	0.14	0.17	0.12
qm_{p11}	0.25	-	0.14	0.14	0.20	0.20
qm_{p12}	0.08	0.14	-	0.08	0.19	0.15
x_{p2}	0.14	0.14	0.08	-	0.21	0.12
qm_{p21}	0.17	0.20	0.19	0.21	-	0.38
qm_{p22}	0.12	0.20	0.15	0.12	0.38	-

Table 5.12: Mutual information estimates between components of the evidence vector, fingerprint, development set.

A						
	x_{p1}	qm_{p11}	qm_{p12}	x_{p2}	qm_{p21}	qm_{p22}
x_{p1}	-	0.20	0.06	0.12	0.06	0.06
qm_{p11}	0.20	-	0.09	0.05	0.03	0.04
qm_{p12}	0.06	0.09	-	0.08	0.07	0.15
x_{p2}	0.12	0.05	0.08	-	0.37	0.09
qm_{p21}	0.06	0.03	0.07	0.37	-	0.11
qm_{p22}	0.06	0.04	0.15	0.09	0.11	-
B						
	x_{p1}	qm_{p11}	qm_{p12}	x_{p2}	qm_{p21}	qm_{p22}
x_{p1}	-	0.39	0.09	0.32	0.27	0.22
qm_{p11}	0.39	-	0.15	0.24	0.28	0.19
qm_{p12}	0.09	0.15	-	0.13	0.15	0.10
x_{p2}	0.32	0.24	0.13	-	0.41	0.23
qm_{p21}	0.27	0.28	0.15	0.41	-	0.37
qm_{p22}	0.22	0.19	0.10	0.23	0.37	-
A+B						
	x_{p1}	qm_{p11}	qm_{p12}	x_{p2}	qm_{p21}	qm_{p22}
x_{p1}	-	0.06	0.02	0.60	0.02	0.02
qm_{p11}	0.06	-	0.08	0.02	0.04	0.04
qm_{p12}	0.02	0.08	-	0.02	0.04	0.07
x_{p2}	0.60	0.02	0.02	-	0.04	0.02
qm_{p21}	0.02	0.04	0.04	0.04	-	0.07
qm_{p22}	0.02	0.04	0.07	0.02	0.07	-
$\overline{(A,B)}$						
	x_{p1}	qm_{p11}	qm_{p12}	x_{p2}	qm_{p21}	qm_{p22}
x_{p1}	-	0.29	0.08	0.22	0.16	0.14
qm_{p11}	0.29	-	0.12	0.15	0.16	0.11
qm_{p12}	0.08	0.12	-	0.10	0.11	0.12
x_{p2}	0.22	0.15	0.10	-	0.39	0.16
qm_{p21}	0.16	0.16	0.11	0.39	-	0.24
qm_{p22}	0.14	0.11	0.12	0.16	0.24	-

Table 5.13: Mutual information estimates between components of the evidence vector, fingerprint, evaluation set.

equivalent to trained multimodal fusion of scores x_{p1} and x_{p2} [16, 83]. In these figures both baseline thresholds are drawn, defining the loci of observations where classification accuracy clearly benefits from the presence of additional evidence dimension. The results of classification in the evidence space defined by $\mathbf{e} = [x_{p1}, x_{p2}]$ is considered baseline for the $Q - stack$ systems applied to multimodal classifier fusion with quality measures for the two dependent fingerprint modalities (optical and thermal sensor).

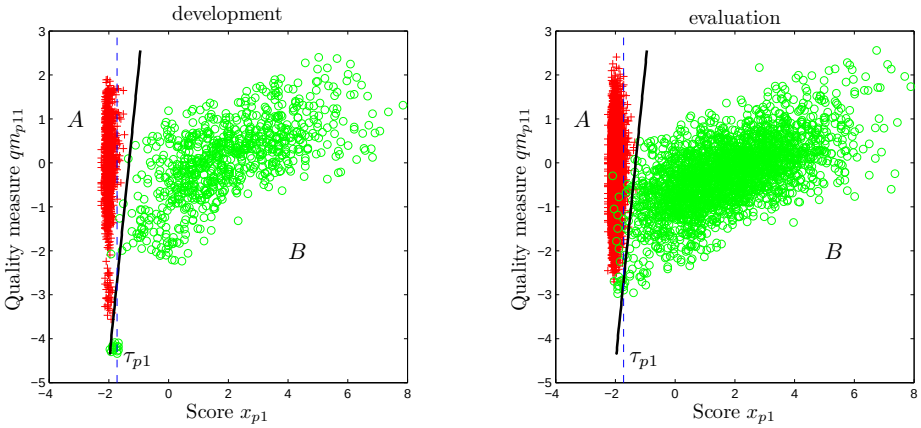


Figure 5.27: Classification in the evidence space using the $SVM - lin$ stacked classifier, fingerprint modality, joint evidence distributions $\mathbf{e} = [x_{p1}, qm_{p11}]$.

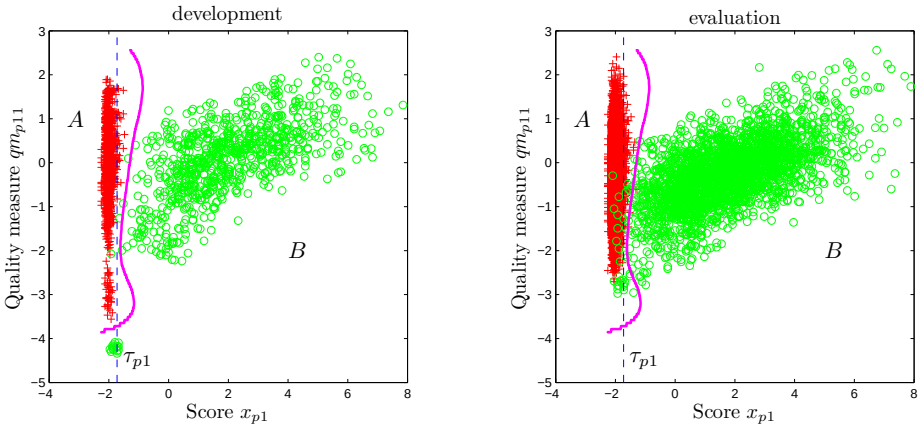


Figure 5.28: Classification in the evidence space using the $SVM - rbf$ stacked classifier, fingerprint modality, joint evidence distributions $\mathbf{e} = [x_{p1}, qm_{p11}]$.

Table 5.14 gives the classification results for the fingerprint modality in terms of total accuracy AC, $HTER$, and error rates per class ER_A and ER_B, for the development data set. Table 5.15 gives the classification results in the same terms for the evaluation data set. In both Tables 5.14 and 5.15 the classification results for all three considered stacked classifiers, $SVM - lin$, $SVM - rbf$ and $Bayes$, are listed.

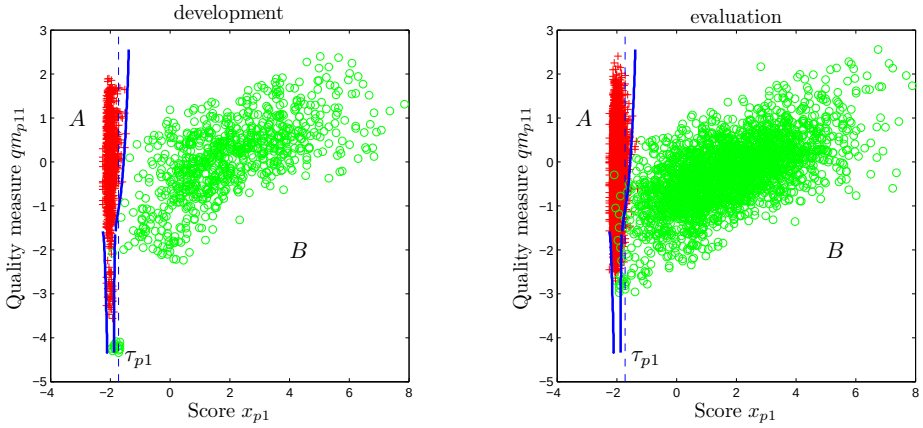


Figure 5.29: Classification in the evidence space using the *Bayes* stacked classifier, fingerprint modality, joint evidence distributions $\mathbf{e} = [x_{p1}, qm_{p11}]$.

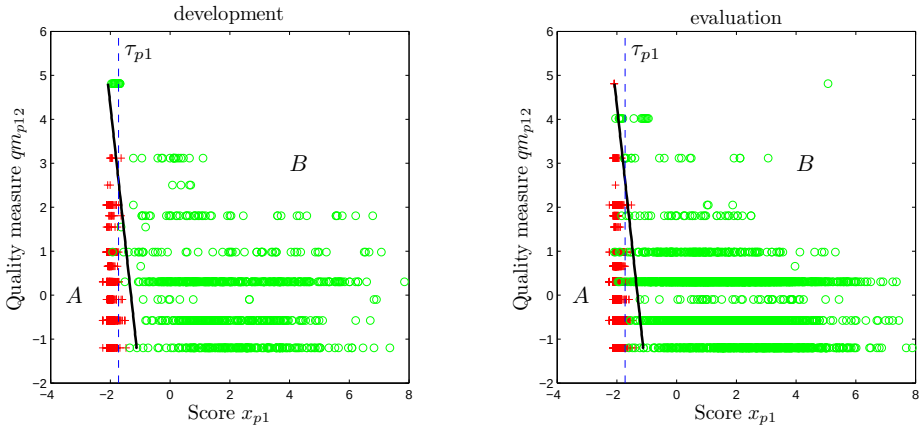


Figure 5.30: Classification in the evidence space using the $SVM - lin$ stacked classifier, fingerprint modality, joint evidence distributions $\mathbf{e} = [x_{p1}, qm_{p12}]$.

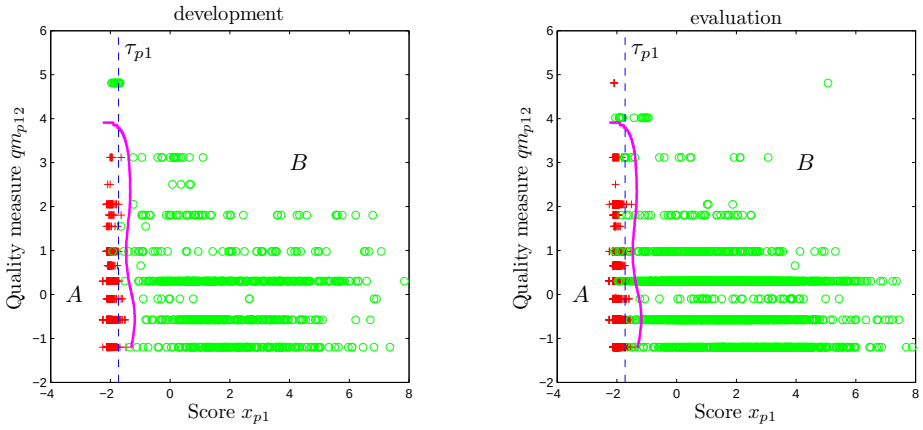


Figure 5.31: Classification in the evidence space using the $SVM - rbf$ stacked classifier, fingerprint modality, joint evidence distributions $\mathbf{e} = [x_{p1}, qm_{p12}]$.

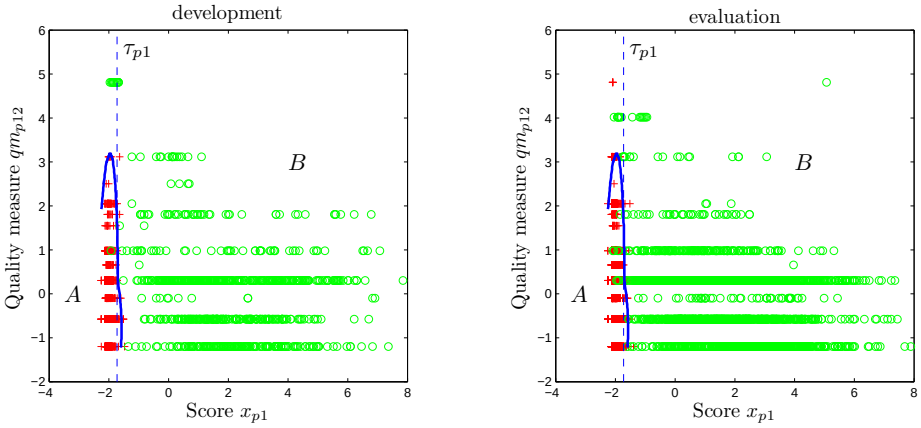


Figure 5.32: Classification in the evidence space using the *Bayes* stacked classifier, fingerprint modality, joint evidence distributions $\mathbf{e} = [x_{p1}, qm_{p12}]$.

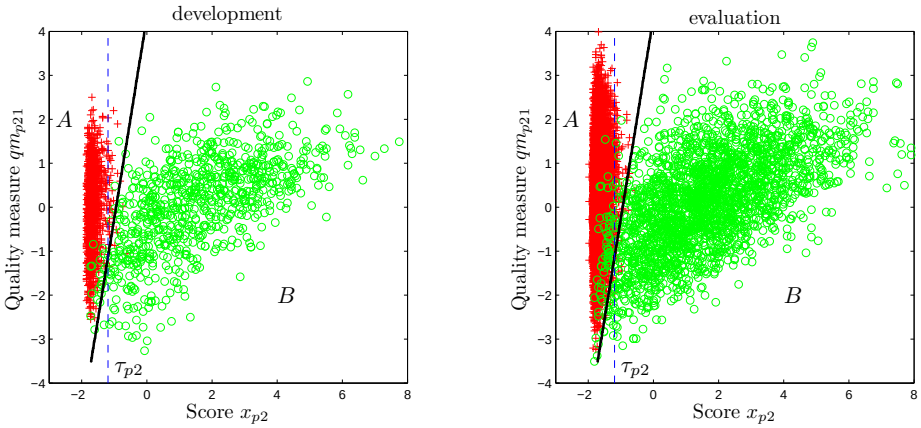


Figure 5.33: Classification in the evidence space using the $SVM - lin$ stacked classifier, fingerprint modality, joint evidence distributions $\mathbf{e} = [x_{p2}, qm_{p21}]$.

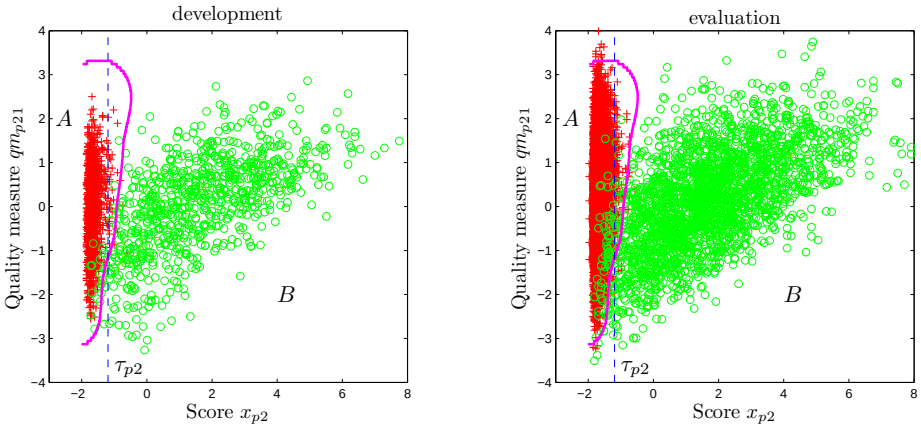


Figure 5.34: Classification in the evidence space using the $SVM - rbf$ stacked classifier, fingerprint modality, joint evidence distributions $\mathbf{e} = [x_{p2}, qm_{p21}]$.

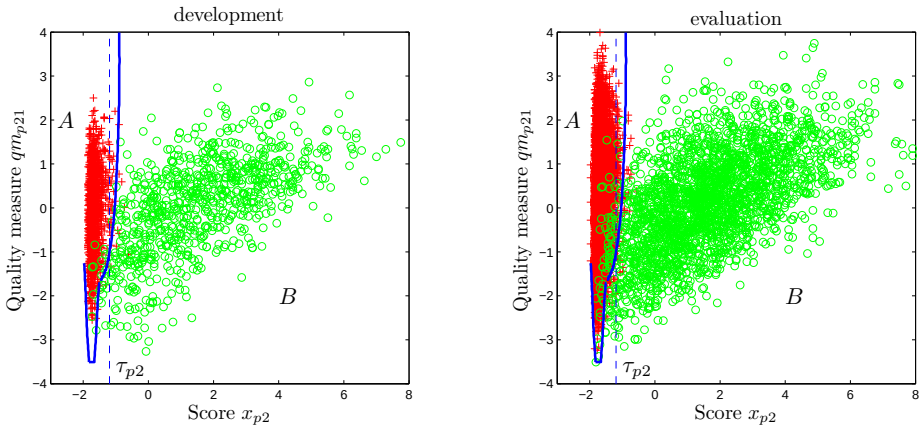


Figure 5.35: Classification in the evidence space using the *Bayes* stacked classifier, fingerprint modality, joint evidence distributions $\mathbf{e} = [x_{p2}, qm_{p21}]$.

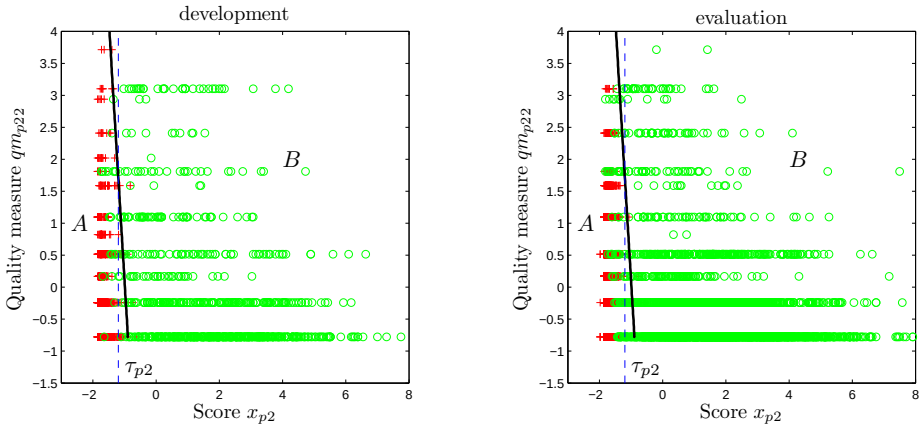


Figure 5.36: Classification in the evidence space using the $SVM - lin$ stacked classifier, fingerprint modality, joint evidence distributions $\mathbf{e} = [x_{p2}, qm_{p22}]$.

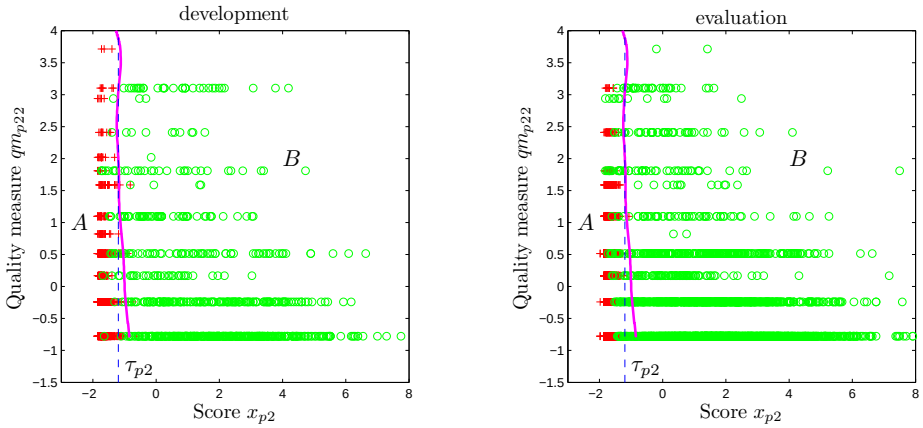


Figure 5.37: Classification in the evidence space using the $SVM - rbf$ stacked classifier, fingerprint modality, joint evidence distributions $\mathbf{e} = [x_{p2}, qm_{p22}]$.

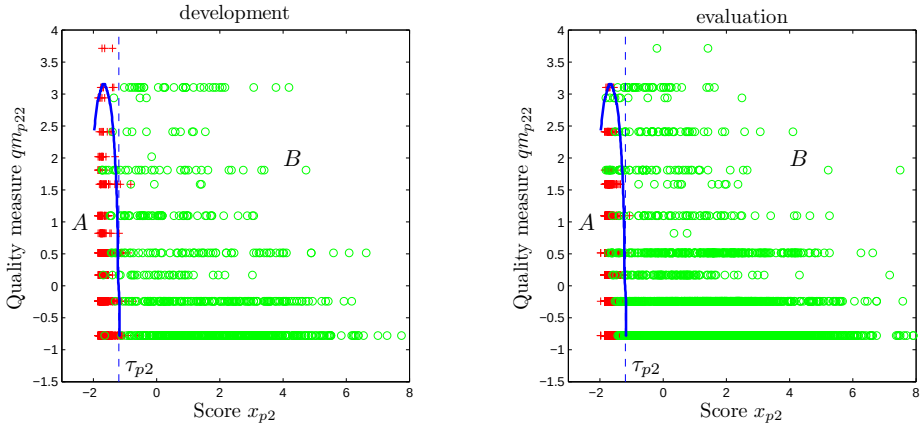


Figure 5.38: Classification in the evidence space using the *Bayes* stacked classifier, fingerprint modality, joint evidence distributions $\mathbf{e} = [x_{p2}, qm_{p22}]$.

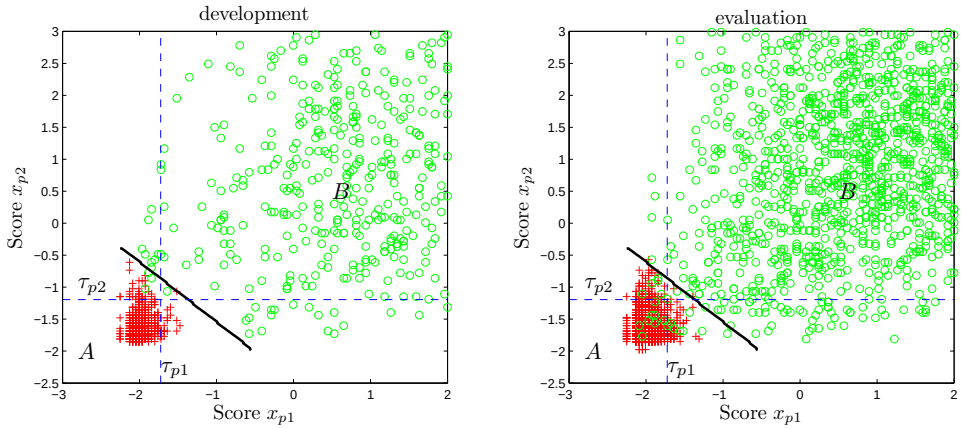


Figure 5.39: Classification in the evidence space using the $SVM - lin$ stacked classifier, fingerprint modality, joint evidence distributions $\mathbf{e} = [x_{p1}, x_{p2}]$.

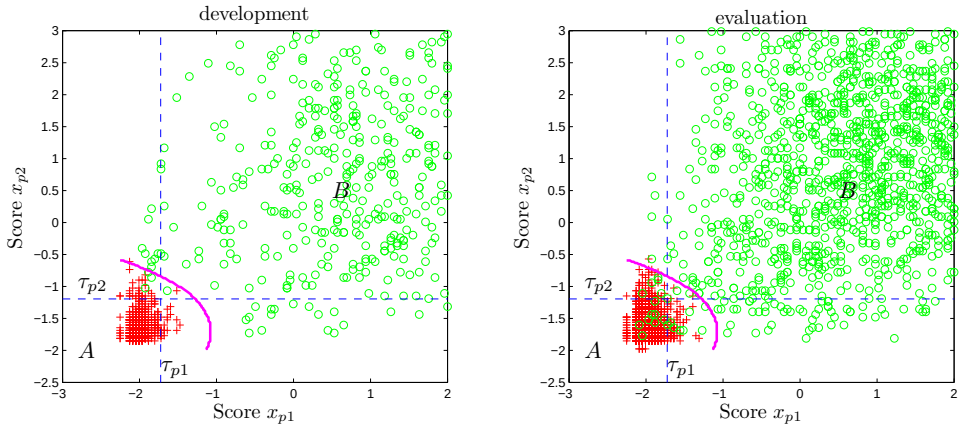


Figure 5.40: Classification in the evidence space using the $SVM - rbf$ stacked classifier, fingerprint modality, joint evidence distributions $\mathbf{e} = [x_{p1}, x_{p2}]$.

	AC	$HTER$	ER_A	ER_B
baseline				
$\mathbf{e} = [x_{p1}]$	*0.9827*	*0.0167*	*0.0138*	*0.0196*
$\mathbf{e} = [x_{p2}]$	*0.9709*	*0.0286*	*0.0263*	*0.0310*
$SVM - lin$				
$\mathbf{e} = [x_{p1}, qm_{p11}]$	0.9975	0.0031	0.0062	0.0000
$\mathbf{e} = [x_{p1}, qm_{p12}]$	0.9965	0.0039	0.0062	0.0016
$\mathbf{e} = [x_{p1}, qm_{p11}, qm_{p12}]$	0.9970	0.0033	0.0050	0.0016
$\mathbf{e} = [x_{p2}, qm_{p21}]$	0.9807	0.0177	0.0100	0.0253
$\mathbf{e} = [x_{p2}, qm_{p22}]$	0.9807	0.0209	0.0288	0.0131
$\mathbf{e} = [x_{p2}, qm_{p21}, qm_{p22}]$	0.9793	0.0189	0.0100	0.0278
$\mathbf{e} = [x_{p1}, x_{p2}]$	*0.9980*	*0.0025*	*0.0050*	*0.0000*
$\mathbf{e} = [x_{p1}, x_{p2}, qm_{p11}, qm_{p21}]$	1.0000	0.0000	0.0000	0.0000
$\mathbf{e} = [x_{p1}, x_{p2}, qm_{p12}, qm_{p22}]$	0.9995	0.0004	0.0000	0.0008
$\mathbf{e} = [x_{p1}, x_{p2}, qm_{p11}, qm_{p21}, qm_{p12}, qm_{p22}]$	1.0000	0.0000	0.0000	0.0000
$SVM - rbf$				
$\mathbf{e} = [x_{p1}, qm_{p11}]$	0.9985	0.0014	0.0012	0.0016
$\mathbf{e} = [x_{p1}, qm_{p12}]$	0.9985	0.0014	0.0012	0.0016
$\mathbf{e} = [x_{p1}, qm_{p11}, qm_{p12}]$	0.9985	0.0014	0.0012	0.0016
$\mathbf{e} = [x_{p2}, qm_{p21}]$	0.9872	0.0126	0.0112	0.0139
$\mathbf{e} = [x_{p2}, qm_{p22}]$	0.9768	0.0231	0.0225	0.0237
$\mathbf{e} = [x_{p2}, qm_{p21}, qm_{p22}]$	0.9891	0.0107	0.0100	0.0114
$\mathbf{e} = [x_{p1}, x_{p2}]$	*0.9970*	*0.0029*	*0.0025*	*0.0033*
$\mathbf{e} = [x_{p1}, x_{p2}, qm_{p11}, qm_{p21}]$	1.0000	0.0000	0.0000	0.0000
$\mathbf{e} = [x_{p1}, x_{p2}, qm_{p12}, qm_{p22}]$	1.0000	0.0000	0.0000	0.0000
$\mathbf{e} = [x_{p1}, x_{p2}, qm_{p11}, qm_{p21}, qm_{p12}, qm_{p22}]$	1.0000	0.0000	0.0000	0.0000
$Bayes$				
$\mathbf{e} = [x_{p1}, qm_{p11}]$	0.9960	0.0039	0.0038	0.0041
$\mathbf{e} = [x_{p1}, qm_{p12}]$	0.9960	0.0037	0.0025	0.0049
$\mathbf{e} = [x_{p1}, qm_{p11}, qm_{p12}]$	0.9985	0.0014	0.0012	0.0016
$\mathbf{e} = [x_{p2}, qm_{p21}]$	0.9867	0.0134	0.0138	0.0131
$\mathbf{e} = [x_{p2}, qm_{p22}]$	0.9728	0.0270	0.0262	0.0278
$\mathbf{e} = [x_{p2}, qm_{p21}, qm_{p22}]$	0.9812	0.0190	0.0200	0.0180
$\mathbf{e} = [x_{p1}, x_{p2}]$	*0.9975*	*0.0025*	*0.0025*	*0.0024*
$\mathbf{e} = [x_{p1}, x_{p2}, qm_{p11}, qm_{p21}]$	0.9995	0.0004	0.0000	0.0008
$\mathbf{e} = [x_{p1}, x_{p2}, qm_{p12}, qm_{p22}]$	0.9970	0.0024	0.0000	0.0049
$\mathbf{e} = [x_{p1}, x_{p2}, qm_{p11}, qm_{p21}, qm_{p12}, qm_{p22}]$	0.9985	0.0012	0.0000	0.0024

Table 5.14: $Q - stack$ results for fingerprint modality, development data set. The results obtained for the baseline systems are marked in italic font.

	AC	$HTER$	ER_A	ER_B
baseline				
$\mathbf{e} = [x_{p1}]$	*0.99042*	*0.008598*	*0.007083*	*0.010112*
$\mathbf{e} = [x_{p2}]$	*0.9849*	*0.021769*	*0.032083*	*0.011454*
$SVM - lin$				
$\mathbf{e} = [x_{p1}, qm_{p11}]$	**0.9903**	**0.007**	**0.0029**	**0.0111**
$\mathbf{e} = [x_{p1}, qm_{p12}]$	**0.9963**	**0.0047**	**0.0062**	**0.0031**
$\mathbf{e} = [x_{p1}, qm_{p11}, qm_{p12}]$	**0.9962**	**0.0039**	**0.0042**	**0.0037**
$\mathbf{e} = [x_{p2}, qm_{p21}]$	0.969	0.0221	0.0083	0.0359
$\mathbf{e} = [x_{p2}, qm_{p22}]$	**0.9878**	**0.0174**	**0.0254**	**0.0093**
$\mathbf{e} = [x_{p2}, qm_{p21}, qm_{p22}]$	0.9677	0.0229	0.0083	0.0374
$\mathbf{e} = [x_{p1}, x_{p2}]$	*0.994*	*0.0059*	*0.0058*	*0.006*
$\mathbf{e} = [x_{p1}, x_{p2}, qm_{p11}, qm_{p21}]$	**0.9976**	**0.0026**	**0.0029**	**0.0023**
$\mathbf{e} = [x_{p1}, x_{p2}, qm_{p12}, qm_{p22}]$	**0.9989**	**0.0025**	**0.0046**	**0.0004**
$\mathbf{e} = [x_{p1}, x_{p2}, qm_{p11}, qm_{p21}, qm_{p12}, qm_{p22}]$	**0.9993**	**0.0019**	**0.0038**	**0**
$SVM - rbf$				
$\mathbf{e} = [x_{p1}, qm_{p11}]$	**0.9926**	**0.0071**	**0.0067**	**0.0075**
$\mathbf{e} = [x_{p1}, qm_{p12}]$	**0.9945**	**0.0051**	**0.0046**	**0.0056**
$\mathbf{e} = [x_{p1}, qm_{p11}, qm_{p12}]$	**0.9931**	**0.0066**	**0.0062**	**0.007**
$\mathbf{e} = [x_{p2}, qm_{p21}]$	0.9783	0.0212	0.0204	0.022
$\mathbf{e} = [x_{p2}, qm_{p22}]$	**0.9843**	**0.0187**	**0.0233**	**0.014**
$\mathbf{e} = [x_{p2}, qm_{p21}, qm_{p22}]$	0.978	0.0229	0.0242	0.0216
$\mathbf{e} = [x_{p1}, x_{p2}]$	*0.9952*	*0.0052*	*0.0058*	*0.0046*
$\mathbf{e} = [x_{p1}, x_{p2}, qm_{p11}, qm_{p21}]$	0.9931	0.0068	0.0067	0.0069
$\mathbf{e} = [x_{p1}, x_{p2}, qm_{p12}, qm_{p22}]$	**0.9961**	**0.0043**	**0.005**	**0.0037**
$\mathbf{e} = [x_{p1}, x_{p2}, qm_{p11}, qm_{p21}, qm_{p12}, qm_{p22}]$	0.9884	0.0118	0.0121	0.0115
$Bayes$				
$\mathbf{e} = [x_{p1}, qm_{p11}]$	**0.9947**	**0.0052**	**0.005**	**0.0054**
$\mathbf{e} = [x_{p1}, qm_{p12}]$	**0.9976**	**0.0039**	**0.0062**	**0.0015**
$\mathbf{e} = [x_{p1}, qm_{p11}, qm_{p12}]$	**0.9978**	**0.003**	**0.0042**	**0.0018**
$\mathbf{e} = [x_{p2}, qm_{p21}]$	**0.9904**	**0.0168**	**0.0279**	**0.0056**
$\mathbf{e} = [x_{p2}, qm_{p22}]$	**0.9899**	**0.0191**	**0.0329**	**0.0052**
$\mathbf{e} = [x_{p2}, qm_{p21}, qm_{p22}]$	**0.9881**	**0.0157**	**0.0217**	**0.0098**
$\mathbf{e} = [x_{p1}, x_{p2}]$	*0.9975*	*0.0046*	*0.0079*	*0.0013*
$\mathbf{e} = [x_{p1}, x_{p2}, qm_{p11}, qm_{p21}]$	**0.9993**	**0.0019**	**0.0038**	**0.0001**
$\mathbf{e} = [x_{p1}, x_{p2}, qm_{p12}, qm_{p22}]$	**0.9976**	**0.0034**	**0.005**	**0.0019**
$\mathbf{e} = [x_{p1}, x_{p2}, qm_{p11}, qm_{p21}, qm_{p12}, qm_{p22}]$	**0.9981**	**0.0026**	**0.0038**	**0.0015**

Table 5.15: $Q - stack$ results for fingerprint modality, evaluation data set. The results obtained for the baseline systems are marked in italic font. The results for classification with quality measures in the proposed $Q - stack$ scheme are in bold font.

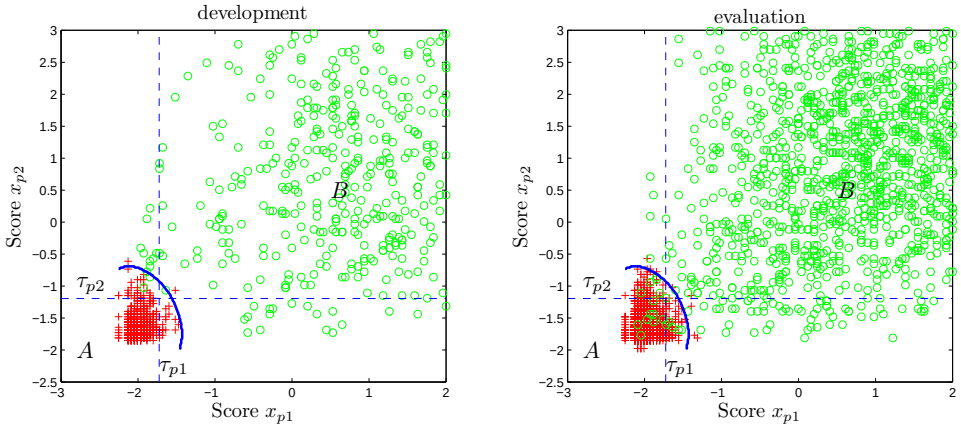


Figure 5.41: Classification in the evidence space using the *Bayes* stacked classifier, fingerprint modality, joint evidence distributions $\mathbf{e} = [x_{p1}, x_{p2}]$.

Discussion of the results

The fingerprint modality belongs to the most accurate and reliable in biometric identity verification, and this fact finds its reflection in particularly low error rates for the fingerprint modality obtained by the baseline systems used in the experiments reported in this Section. The already very good baseline classifier performance makes the task of improving classification and lowering error rates particularly hard. Yet the proposed method of $Q - stack$ lived up to our expectations and offered classification performance superior to the baseline systems.

The within-class estimates of dependence between scores and quality measures are well pronounced, in particular between x_{p1} and qm_{p11}, as well as between x_{p2} and qm_{p21}. In general, both dependence estimates hint on the fact that the quality measures qm_{p11} and qm_{p21} sport significantly stronger dependencies on the similarity scores than it is the case of qm_{p12} and qm_{p22}. This is noteworthy since qm_{p12} and qm_{p22}, the NFIQ quality measures native to the NFIS2 software package used as matcher here, were designed and conceived with the purpose of predicting the classifier scores. However, unlike the frequency-domain quality measure qm_{p11} and qm_{p21} the NFIQ quality measures qm_{p12} and qm_{p22} are discrete, which is bound to cause the correlation coefficient and mutual information estimates to be lower than they would be in the continuous case. Despite the relatively low dependence estimates between the NFIQ quality measures and the classification scores, the application of $Q - stack$ to classification of evidence that contains qm_{p12} and qm_{p22} was successful in terms of achieving accuracy gains over baseline classifiers.

The existence of dependencies between the scores and quality measures can be detected visually from the graphs shown in figures from 5.29 to 5.37. If the within-class dependencies between the scores and quality measures exist, the joint class-conditional evidence distributions are not positioned along either of the main axes of the coordinate system. Instead, these distributions are slanted with respect to the axes. A very good example of such distribution slant due to the evidence dependencies is shown in Figure 5.27, where the distribution of $\mathbf{e}|B$ is clearly and visibly 'slanted', while $\mathbf{e}|A$ is distributed mostly along the vertical axis of quality measure values. The reader shall discover similar spatial configurations of class-conditional evidence distributions in other figures in this section and in this chapter. These configurations, which reveal the presence of dependencies between the evidence components, lead to improved class separation, which a properly chosen $Q - stack$ classifier can exploit. This effect finds its reflection in the classification results shown in Table 5.15.

Classification results gathered in Table 5.15 show that both the $SVM - lin$ and the *Bayes* classifiers were capable of taking advantage of quality measures included in the evidence, for the overwhelming majority of considered configurations of the evidence vector. Similarly as in the case of the face modality, the $SVM - rbf$ classifier proved to overfit the decision boundary to the training data as a result of the insufficient development data sample size to represent the population properly. Simplifying assumptions and complexity limitations put on the $SVM - lin$ and the *Bayes* classifiers helped overcome this problem and delivered classification accuracy superior of that of the baseline systems, for both single-classifier and multimodal scenarios.

5.3.4 Error prediction and rejection

In this section we demonstrate the credence-based error prediction and rejection technique presented in Chapter 4, applied to fingerprint modality. We consider here data originating from both available fingerprint modalities, acquired using the optical and thermal scanners, and their multimodal fusion. For every modality and for their multimodal combination, we present results for credence-based error prediction and rejection obtained with and without the use of relevant quality measures. The aim of the experiment is to demonstrate that for the fingerprint modality an accurate error prediction can be performed based on credence estimate method presented in chapter 4, and that the method can be applied when quality measure are available as well as when they are not.

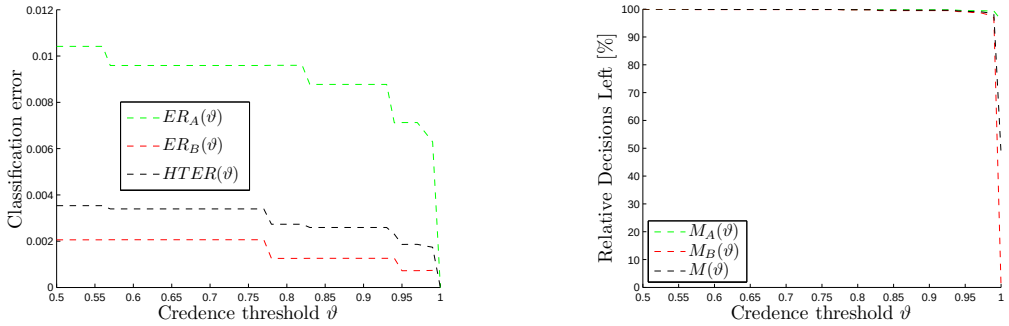
For the fingerprint modality, following evidence combinations were used in the error prediction experiments:

- **Single classifier, optical scanner:** $\mathbf{e} = [x_{p1}]$, $\mathbf{e} = [x_{p1}, qm_{p11}, qm_{p12}]$. The classification error as a function of the credence threshold is shown in Figures 5.42(a) and 5.43(a), respectively. The percentage of decisions remaining after credence thresholding as a function of the credence threshold is given in Figures 5.42(b) and 5.43(b), respectively.
- **Single classifier, thermal scanner:** $\mathbf{e} = [x_{p2}]$, $\mathbf{e} = [x_{p2}, qm_{p21}, qm_{p22}]$. The classification error as a function of the credence threshold is shown in Figures 5.44(a) and 5.45(a), respectively. The percentage of decisions remaining after credence thresholding as a function of the credence threshold is given in Figures 5.44(b) and 5.45(b), respectively.
- **Multi-modal fusion, optical + thermal scanners:** $\mathbf{e} = [x_{p1}, x_{p2}]$, $\mathbf{e} = [x_{p1}, x_{p2}, qm_{p11}, qm_{p12}, qm_{p21}, qm_{p22}]$. The classification error as a function of the credence threshold is shown in Figures 5.46(a) and 5.47(a), respectively. The percentage of decisions remaining after credence thresholding as a function of the credence threshold is given in Figures 5.46(b) and 5.47(b), respectively.

Numerical results gathered in Table 5.16 allow for an the evaluation of the credence estimators according to the accountability criterion. For each considered configuration of evidence, the mean observed total classification accuracy AC, Half-Total Error Rate $HTER$ and per-class errors ER_A and ER_B are reported together with their standard deviations after 100 experimental runs. Corresponding mean values of credence estimators with their standard deviations are given. The differences between the mean observed values and the mean credence predictors are reported (*Mean difference* δ). According to the accountability criterion, the smaller these differences the more accurate the error prediction, and consequently the better the credence estimator.

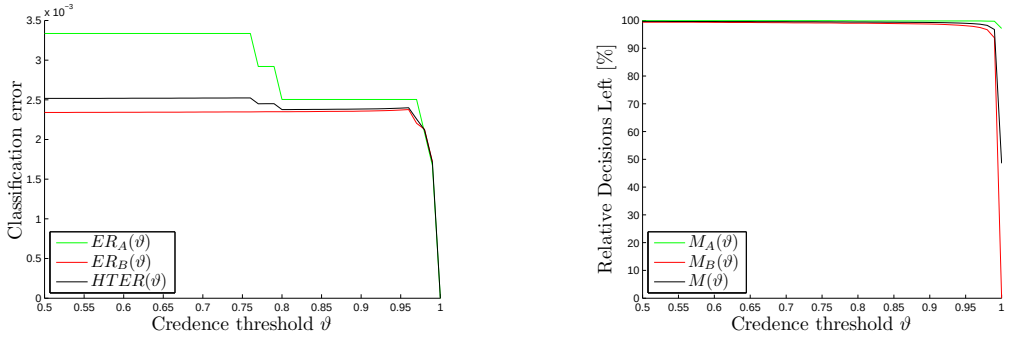
Discussion of the results

Figures 5.42, 5.43, 5.44, 5.45, 5.46 and 5.47 show that classification error expressed in terms of $HTER$, ER_A and ER_B decreases with the growing value of the credence threshold ϑ for all consid-



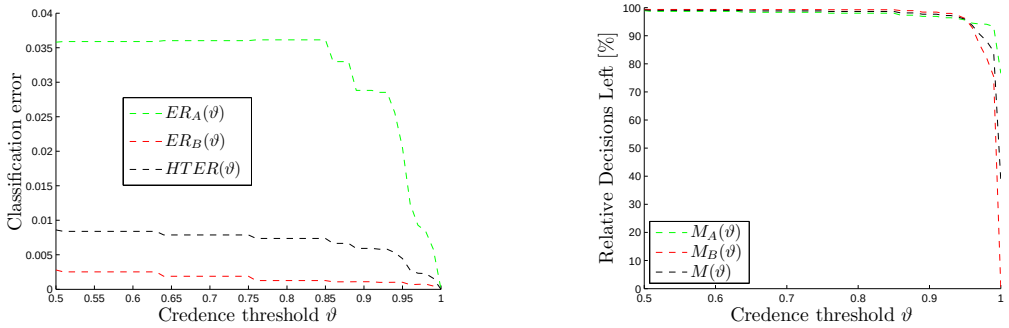
(a) error after credence thresholding (b) percentage of decisions after credence thresholding

Figure 5.42: Error prediction and rejection using credence estimates, fingerprint, optical sensor, evidence configuration $\mathbf{e} = [x_{p1}]$.



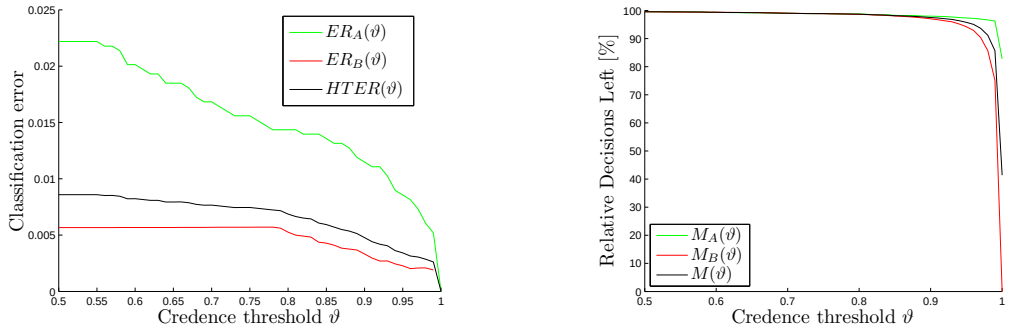
(a) error after credence thresholding (b) percentage of decisions after credence thresholding

Figure 5.43: Error prediction and rejection using credence estimates, fingerprint, optical sensor, evidence configuration $\mathbf{e} = [x_{p1}, qm_{p11}, qm_{p12}]$.



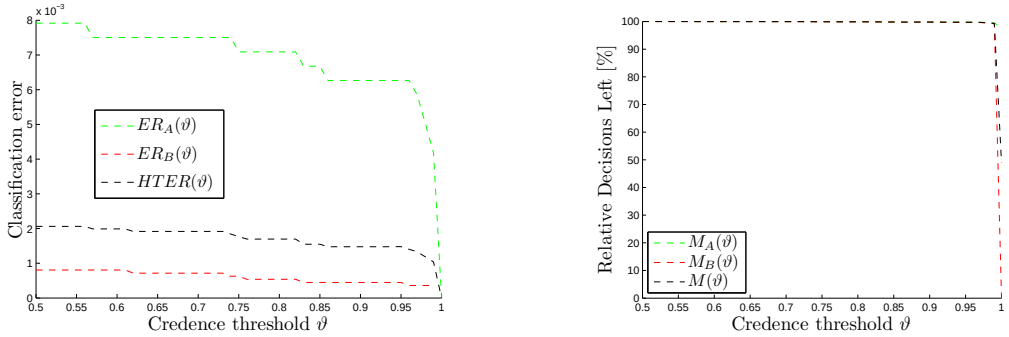
(a) error after credence thresholding (b) percentage of decisions after credence thresholding

Figure 5.44: Error prediction and rejection using credence estimates, fingerprint, thermal sensor, evidence configuration $\mathbf{e} = [x_{p2}]$.



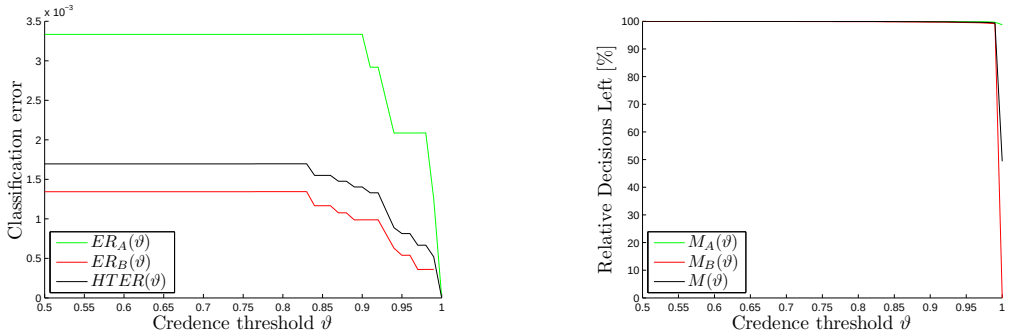
(a) error after credence thresholding (b) percentage of decisions after credence thresholding

Figure 5.45: Error prediction and rejection using credence estimates, fingerprint, thermal sensor, evidence configuration $\mathbf{e} = [x_{p2}, qm_{p21}, qm_{p22}]$.



(a) error after credence thresholding (b) percentage of decisions after credence thresholding

Figure 5.46: Error prediction and rejection using credence estimates, multimodal fingerprint fusion, optical + thermal sensors, evidence configuration $\mathbf{e} = [x_{p1}, x_{p2}]$.



(a) error after credence thresholding (b) percentage of decisions after credence thresholding

Figure 5.47: Error prediction and rejection using credence estimates, multimodal fingerprint fusion, optical + thermal sensors, evidence configuration $\mathbf{e} = [x_{p1}, x_{p2}, qm_{p11}, qm_{p12}, qm_{p21}, qm_{p22}]$.

	AC	$HTER$	ER_A	ER_B
$\mathbf{e} = [x_{p1}]$				
Actual accuracy	0.9908	0.0090	0.0088	0.0093
St. dev. of actual accuracy	0.0000	0.0000	0.0000	0.0000
Credence estimate	0.9913	0.0061	0.0101	0.0021
St. dev. of credence estimate	0.0005	0.0006	0.0005	0.0006
Mean difference δ	**-0.0005**	**0.0029**	**-0.0013**	**0.0072**
$\mathbf{e} = [x_{p1}, qm_{p11}, qm_{p12}]$				
Actual accuracy	0.9946	0.0052	0.0048	0.0055
St. dev. of actual accuracy	0.0001	0.0001	0.0002	0.0001
Credence estimate	0.9961	0.0029	0.0045	0.0014
St. dev. of credence estimate	0.0001	0.0001	0.0001	0.0001
Mean difference δ	**-0.0014**	**0.0022**	**0.0003**	**0.0040**
$\mathbf{e} = [x_{p2}]$				
Actual accuracy	0.9863	0.0222	0.0354	0.0090
St. dev. of actual accuracy	0.0000	0.0000	0.0000	0.0000
Credence estimate	0.9780	0.0184	0.0240	0.0127
St. dev. of credence estimate	0.0010	0.0010	0.0010	0.0011
Mean difference	**0.0083**	**0.0038**	**0.0114**	**-0.0037**
$\mathbf{e} = [x_{p2}, qm_{p21}, qm_{p22}]$				
Actual accuracy	0.9904	0.0168	0.0279	0.0056
St. dev. of actual accuracy	0.0000	0.0000	0.0000	0.0000
Credence estimate	0.9889	0.0099	0.0118	0.0079
St. dev. of credence estimate	0.0000	0.0000	0.0000	0.0000
Mean difference δ	**0.0015**	**0.0069**	**0.0161**	**-0.0023**
$\mathbf{e} = [x_{p1}, x_{p2}]$				
Actual accuracy	0.9975	0.0046	0.0079	0.0013
St. dev. of actual accuracy	0.0000	0.0001	0.0001	0.0000
Credence estimate	0.9988	0.0010	0.0013	0.0006
St. dev. of credence estimate	0.0000	0.0000	0.0000	0.0001
Mean difference δ	**-0.0013**	**0.0036**	**0.0066**	**0.0007**
$\mathbf{e} = [x_{p1}, x_{p2}, qm_{p11}, qm_{p12}, qm_{p21}, qm_{p22}]$				
Actual accuracy	0.9993	0.0019	0.0038	0.0001
St. dev. of actual accuracy	0.0000	0.0000	0.0000	0.0000
Credence estimate	0.9990	0.0010	0.0010	0.0009
St. dev. of credence estimate	0.0000	0.0000	0.0000	0.0000
Mean difference δ	**0.0003**	**0.0009**	**0.0028**	**-0.0008**

Table 5.16: Evaluation of credence estimates using the accountability criterion, fingerprint modality, after 100 experimental iterations

ered evidence configurations. This behavior is in accord with the theoretical expectations discussed in Section 4.2.1, and it agrees with the intuitive notion of the concept of decision credence: discarding classification decisions deemed not reliable in terms of its credence estimate indeed resulted in reduced observed error over the remaining decisions. This effect shows that the relative number of erroneous decisions in the discarded pool of decisions was higher than that in the decisions considered credible according to the credence estimates. In all presented graphs $HTER$ decreases monotonically as ϑ grows, which satisfies the monotonicity criterion for evaluation of the credence estimates.

The accuracy of the credence estimates for the fingerprint experiments in terms of the accountability criterion can be gauged by the values of δ given by Table 5.16. The accuracy of obtained mean credence estimates is high in comparison with the classification accuracies, which allows for a conclusion that the accountability criterion is met. Note that the observed very low error rates obtained on a evaluation data set of a limited size cause that a change of class label of a single classified sample entails proportionally large change in the observed error rate (Table 5.2). The same holds for the mean credence estimates and therefore it is not meaningful to attempt a statistical significance analysis of the presented results.

The presented results show that the presented method of credence estimation allows for accurate error prediction for single-modality fingerprint verification as well as in the multimodal fingerprint verification, in both cases with- and without an involvement of quality measures.

It is particularly worth mentioning that the observed accuracy gain for, already very accurate, fingerprint verification was possible by discarding only a very small number of unreliable decisions, whose credence estimates fell below the credence threshold. Note that for all subfigures (b) the percentage of remaining decisions remains close to 100% until the credence threshold reaches values very close to $\vartheta = 1$. Despite this, the resulting error rates after credence thresholding are reduced for much lower values of $\vartheta \ll 1$. This results shows that the proposed method can be of practical use in applications where high level of confidence in classification decisions is required. In such applications, rejected decision can, for instance, entail a request for a repeated signal acquisition, as proposed in [129].

5.4 Multimodal fusion: face + fingerprint

In Section 5.3 we have presented results for fusion scores for dependent fingerprint modalities, namely for images acquired using the optical and thermal scanners. Since in both acquisition channels data of very similar character was collected, this represents indeed a special case of multimodal fusion. Although this fact does not undermine the validity of presented results concerning the benefits of applying the proposed scheme of $Q - stack$, one may argue that a fusion of highly dependent modalities is not the most common type of fusion seen in the domain of biometric identity verification [134]. For this reason, below we present results of a multimodal biometric fusion in a traditional sense, namely a fusion of the face and fingerprint modalities.

In the multimodal fusion experiments reported here we use both available face modality scores - x_{f1},x_{f2}, and both available fingerprint modality scores - x_{p1},x_{p2}, and their associated quality measures. Since each of the scores and quality measures could be placed in the evidence vector, the total number of possible evidence combinations for $Q - stack$ in multimodal fusion is vast, and the number of possible experiments to report is tripled due to the fact that we have three different stacked classifiers at our disposal. Such a number of experiments would neither fit the frames of this thesis nor would it help shed more light on the main topic of this work. Therefore the experiments reported below are a selection of fusion experiments involving two score components in the evidence

development					evaluation				
A									
	x_{f1}	x_{f2}	x_{p1}	x_{p2}		x_{f1}	x_{f2}	x_{p1}	x_{p2}
x_{f1}	1.00	0.35	0.03	-0.05	x_{f1}	1.00	0.23	-0.10	-0.08
x_{f2}	0.35	1.00	-0.01	0.01	x_{f2}	0.23	1.00	-0.11	-0.10
x_{p1}	0.03	-0.01	1.00	0.03	x_{p1}	-0.10	-0.11	1.00	0.18
x_{p2}	-0.05	0.01	0.03	1.00	x_{p2}	-0.08	-0.10	0.18	1.00
B									
	x_{f1}	x_{f2}	x_{p1}	x_{p2}		x_{f1}	x_{f2}	x_{p1}	x_{p2}
x_{f1}	1.00	0.14	0.08	0.04	x_{f1}	1.00	0.15	-0.10	0.11
x_{f2}	0.14	1.00	0.24	0.06	x_{f2}	0.15	1.00	-0.06	0.13
x_{p1}	0.08	0.24	1.00	0.36	x_{p1}	-0.10	-0.06	1.00	0.36
x_{p2}	0.04	0.06	0.36	1.00	x_{p2}	0.11	0.13	0.36	1.00
$A+B$									
	x_{f1}	x_{f2}	x_{p1}	x_{p2}		x_{f1}	x_{f2}	x_{p1}	x_{p2}
x_{f1}	1.00	0.55	0.68	0.63	x_{f1}	1.00	0.36	0.60	0.58
x_{f2}	0.55	1.00	0.54	0.48	x_{f2}	0.36	1.00	0.36	0.35
x_{p1}	0.68	0.54	1.00	0.84	x_{p1}	0.60	0.36	1.00	0.84
x_{p2}	0.63	0.48	0.84	1.00	x_{p2}	0.58	0.35	0.84	1.00
$\overline{(A,B)}$									
	x_{f1}	x_{f2}	x_{p1}	x_{p2}		x_{f1}	x_{f2}	x_{p1}	x_{p2}
x_{f1}	1.00	0.25	0.05	0.00	x_{f1}	1.00	0.19	-0.10	0.01
x_{f2}	0.25	1.00	0.11	0.03	x_{f2}	0.19	1.00	-0.09	0.02
x_{p1}	0.05	0.11	1.00	0.19	x_{p1}	-0.10	-0.09	1.00	0.27
x_{p2}	0.00	0.03	0.19	1.00	x_{p2}	0.01	0.02	0.27	1.00

Table 5.17: Pair-wise computed Pearson's correlation coefficients between similarity scores from the face and fingerprint modalities.

vector (one score component from each involved modality), and of experiments involving all available scores from both modalities. We also limit the experiments reported here to $SVM-lin$ classifier, the best-performing stacked classifier in the face- and fingerprint experiments.

Analysis of dependencies between evidence components

In Sections 5.2 and 5.3 we have presented the estimates of dependencies that exist between the components of the evidence vector within the face and the fingerprint modalities. Naturally the same dependencies do transfer to the multimodal experiments described in the current Section. However, previously posted dependence estimates do not quantify the relationships between scores originating from the face modality and those originating from the fingerprint modality. These dependence estimates, important in understanding the structure of the classification problem in the multimodal fusion scenario, are shown in Tables 5.17 (Pearson's correlation coefficient, Equation 5.1) and (mutual information, Equation 5.2).

The relations between different classification scores can be used as a measure of base classifier diversity, which has been shown to be of importance for a successful classifier combination [96]. Classifier diversity together with its accuracy are indicators of the gain that one can expect from utilizing them together in a multi-classifier ensemble: the more accurate and the more diversified

development					evaluation				
A									
	x_{f1}	x_{f2}	x_{p1}	x_{p2}		x_{f1}	x_{f2}	x_{p1}	x_{p2}
x_{f1}	-	0.11	0.01	0.01	x_{f1}	-	0.09	0.06	0.08
x_{f2}	0.11	-	0.01	0.00	x_{f2}	0.09	-	0.08	0.07
x_{p1}	0.01	0.01	-	0.01	x_{p1}	0.06	0.08	-	0.12
x_{p2}	0.01	0.00	0.01	-	x_{p2}	0.08	0.07	0.12	-
B									
	x_{f1}	x_{f2}	x_{p1}	x_{p2}		x_{f1}	x_{f2}	x_{p1}	x_{p2}
x_{f1}	-	0.08	0.11	0.10	x_{f1}	-	0.15	0.12	0.14
x_{f2}	0.08	-	0.13	0.13	x_{f2}	0.15	-	0.17	0.14
x_{p1}	0.11	0.13	-	0.28	x_{p1}	0.12	0.17	-	0.32
x_{p2}	0.10	0.13	0.28	-	x_{p2}	0.14	0.14	0.32	-
$A+B$									
	x_{f1}	x_{f2}	x_{p1}	x_{p2}		x_{f1}	x_{f2}	x_{p1}	x_{p2}
x_{f1}	-	0.30	0.55	0.50	x_{f1}	-	0.15	0.33	0.30
x_{f2}	0.30	-	0.31	0.29	x_{f2}	0.15	-	0.14	0.13
x_{p1}	0.55	0.31	-	0.93	x_{p1}	0.33	0.14	-	0.60
x_{p2}	0.50	0.29	0.93	-	x_{p2}	0.30	0.13	0.60	-
(A,B)									
	x_{f1}	x_{f2}	x_{p1}	x_{p2}		x_{f1}	x_{f2}	x_{p1}	x_{p2}
x_{f1}	-	0.09	0.06	0.05	x_{f1}	-	0.12	0.09	0.11
x_{f2}	0.09	-	0.07	0.07	x_{f2}	0.12	-	0.13	0.11
x_{p1}	0.06	0.07	-	0.14	x_{p1}	0.09	0.13	-	0.22
x_{p2}	0.05	0.07	0.14	-	x_{p2}	0.11	0.11	0.22	-

Table 5.18: Pair-wise mutual information estimates for similarity scores obtained for the face and fingerprint modalities.

the base classifiers are, the more gain out of combining them can be anticipated [108]. However, care must be taken when interpreting the dependencies between the base classifiers as sole straightforward predictors or causes of improved class separation over individual classifiers. As the reader must bear in mind, none of the scores can be considered as a quality measure since all scores par excellence are class-dependent and therefore the effects of dependencies on class separation mentioned in Appendix A apply. Nevertheless, the proposed data-driven approach to gauging the benefits from classification in the evidence space still remains valid.

However, we leave this topic on a sidetrack since analyzing the feature dependencies in generic classifier combination is not necessary for an understanding of the role of class-independent quality measures in classification. Instead, in the following section we will show how introducing these class-independent quality measures into the classifier ensemble helps reduce classification errors, according to the theoretical findings which we spelled out in Section 3.4.4.

Classification in the evidence space

Following figures show the decision boundaries created by the baseline fusion classifiers applied to multimodal fusion of scores obtained for the face and fingerprint classifiers. All pair-wise multimodal combinations of face and fingerprint are considered. In the experiments below we also consider a fusion of all available scores, but due to its dimensionality this classification problem could not be presented graphically. As the baseline fusion classifiers the same three stacked classifiers are used as in the previous Sections: $SVM-lin$, $SVM-rbf$ and $Bayes$. Figures 5.48 to 5.51 prove that the multimodal classifier fusion is indeed a special case of $Q-stack$.

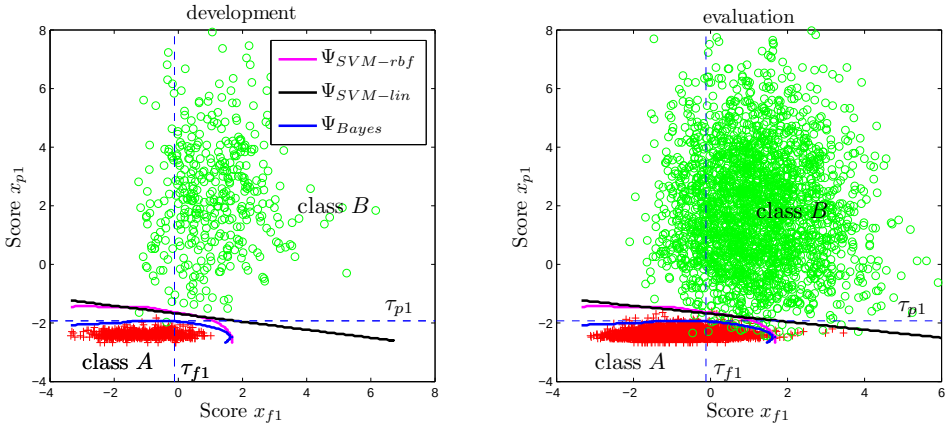


Figure 5.48: Classification in the evidence space using the $SVM-lin$, $SVM-rbf$ and $Bayes$ stacked classifiers, fusion of face and fingerprint modalities, joint evidence distributions $\mathbf{e} = [x_{f1}, x_{p1}]$.

The classification results for baseline multimodal fusion without quality measures are listed in Table 5.19. As we mentioned before, in the multimodal $Q-stack$ experiments we are using only the most competent $SVM-lin$ classifier. This is due to the disproportion between the problem dimensionality and the size of available training data, which led to frequent classifier overtraining in the case of the $SVM-rbf$ and $Bayes$ stacked classifiers in the reported face and fingerprint experiments. Having said that, this gives us no reason to limit the choice of best-performing multimodal fusion baseline to only the $SVM-lin$ classifier. Instead, the best performing fusion classifier is chosen as baseline for comparison with $Q-stack$ results for multimodal fusion. The best-performing

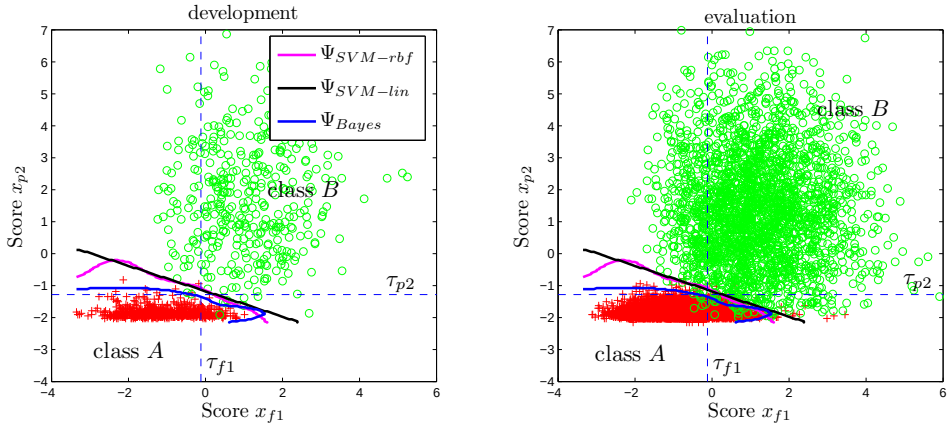


Figure 5.49: Classification in the evidence space using the $SVM - lin$, $SVM - rbf$ and $Bayes$ stacked classifiers, fusion of face and fingerprint modalities, joint evidence distributions $\mathbf{e} = [x_{f1}, x_{p2}]$.

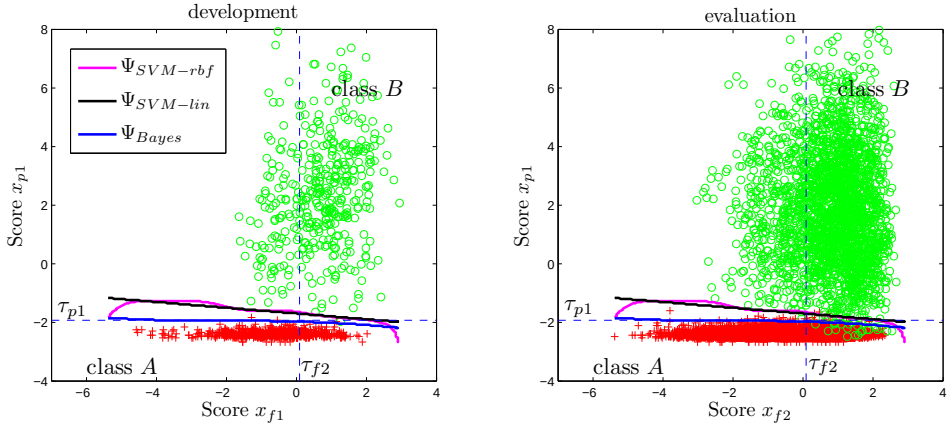


Figure 5.50: Classification in the evidence space using the $SVM - lin$, $SVM - rbf$ and $Bayes$ stacked classifiers, fusion of face and fingerprint modalities, joint evidence distributions $\mathbf{e} = [x_{f2}, x_{p1}]$.

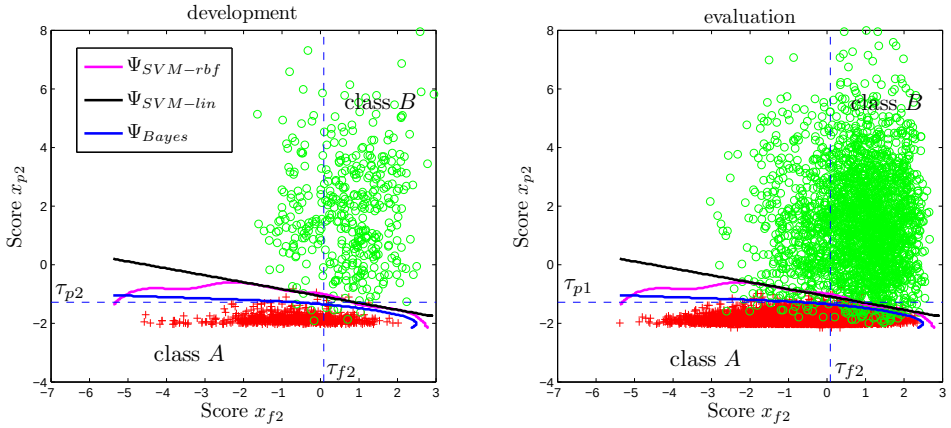


Figure 5.51: Classification in the evidence space using the $SVM-lin$, $SVM-rbf$ and $Bayes$ stacked classifiers, fusion of face and fingerprint modalities, joint evidence distributions $\mathbf{e} = [x_{f2}, x_{p2}]$.

development					evaluation				
face: 1, fingerprint: 1, $\mathbf{e} = [x_{f1}, x_{p1}]$									
$SVM-lin$	0.9991	0.0012	0.0025	0	$SVM-lin$	**0.9969**	**0.0038**	**0.005**	**0.0027**
$SVM-rbf$	0.9973	0.0027	0.0025	0.0028	$SVM-rbf$	0.9957	0.0043	0.0042	0.0044
$Bayes$	0.9973	0.0027	0.0025	0.0028	$Bayes$	0.9954	0.0045	0.0042	0.0047
face: 1, fingerprint: 2, $\mathbf{e} = [x_{f1}, x_{p2}]$									
$SVM-lin$	0.9918	0.0069	0.0025	0.0114	$SVM-lin$	**0.991**	**0.0086**	**0.0079**	**0.0092**
$SVM-rbf$	0.9918	0.008	0.0075	0.0085	$SVM-rbf$	0.9901	0.0098	0.0096	0.01
$Bayes$	0.9927	0.0073	0.0075	0.0071	$Bayes$	0.9920	0.0095	0.0117	0.0072
face: 2, fingerprint: 1, $\mathbf{e} = [x_{f2}, x_{p1}]$									
$SVM-lin$	0.9991	0.0012	0.0025	0	$SVM-lin$	0.9897	0.0076	0.0033	0.0118
$SVM-rbf$	1	0	0	0	$SVM-rbf$	**0.9949**	**0.0049**	**0.0046**	**0.0052**
$Bayes$	0.9982	0.002	0.0025	0.0014	$Bayes$	0.9972	0.0056	0.01	0.0013
face: 2, fingerprint: 2, $\mathbf{e} = [x_{f2}, x_{p2}]$									
$SVM-lin$	0.9837	0.0166	0.0175	0.0156	$SVM-lin$	0.9906	0.0232	0.0446	0.0018
$SVM-rbf$	0.9837	0.0166	0.0175	0.0156	$SVM-rbf$	0.9812	0.0193	0.02	0.0185
$Bayes$	0.9828	0.0173	0.0175	0.0171	$Bayes$	**0.9817**	**0.0189**	**0.02**	**0.0179**
face: 1+2, fingerprint: 1+2, $\mathbf{e} = [x_{f1}, x_{f2}, x_{p1}, x_{p2}]$									
$SVM-lin$	1	0	0	0	$SVM-lin$	**0.9992**	**0.0016**	**0.0029**	**0.0004**
$SVM-rbf$	1	0	0	0	$SVM-rbf$	0.9943	0.0058	0.0058	0.0057
$Bayes$	1	0	0	0	$Bayes$	0.998	0.0024	0.0029	0.0018

Table 5.19: Baseline multimodal fusion results, face + fingerprint, for three fusion classifiers. The best results for each combination of modalities are marked in bold font.

classifiers for each modality combination, chosen based on its $HTER$ obtained on the evaluation set, are marked in bold font in Table 5.19. Note that the $SVM-lin$ classifier again turned out to be on average the most competent of the three considered stacked classifiers.

Tables 5.20 and 5.21 present the results of $Q-stack$ using the $SVM-lin$ as the stacked classifier, for the multimodal fusion experiment. In this experiment, the baseline evidence combinations of face and fingerprint scores have been augmented with different configurations of quality measures, coming

	AC	$HTER$	ER_A	ER_B
face: 1, fingerprint: 1				
$\mathbf{e} = [x_{f1}, x_{p1}, qm_{f1}]$	0.9991	0.0012	0.0025	0
$\mathbf{e} = [x_{f1}, x_{p1}, qm_{p11}]$	1	0	0	0
$\mathbf{e} = [x_{f1}, x_{p1}, qm_{p12}]$	0.9991	0.0012	0.0025	0
$\mathbf{e} = [x_{f1}, x_{p1}, qm_{f1}, qm_{p11}]$	1	0	0	0
$\mathbf{e} = [x_{f1}, x_{p1}, qm_{f1}, qm_{p12}]$	0.9991	0.0012	0.0025	0
$\mathbf{e} = [x_{f1}, x_{p1}, qm_{f1}, qm_{p11}, qm_{p12}]$	**1**	**0**	**0**	**0**
face: 1, fingerprint: 2				
$\mathbf{e} = [x_{f1}, x_{p2}, qm_{f1}]$	0.9955	0.0057	0.01	0.0014
$\mathbf{e} = [x_{f1}, x_{p2}, qm_{p21}]$	0.9973	0.0027	0.0025	0.0028
$\mathbf{e} = [x_{f1}, x_{p2}, qm_{p22}]$	0.9982	0.002	0.0025	0.0014
$\mathbf{e} = [x_{f1}, x_{p2}, qm_{f1}, qm_{p21}]$	0.9982	0.002	0.0025	0.0014
$\mathbf{e} = [x_{f1}, x_{p2}, qm_{f1}, qm_{p22}]$	0.9964	0.0034	0.0025	0.0043
$\mathbf{e} = [x_{f1}, x_{p2}, qm_{f1}, qm_{p21}, qm_{p22}]$	**0.9973**	**0.0027**	**0.0025**	**0.0028**
face: 2, fingerprint: 1				
$\mathbf{e} = [x_{f2}, x_{p1}, qm_{f2}]$	0.9991	0.0007	0	0.0014
$\mathbf{e} = [x_{f2}, x_{p1}, qm_{p11}]$	1	0	0	0
$\mathbf{e} = [x_{f2}, x_{p1}, qm_{p12}]$	0.9991	0.0007	0	0.0014
$\mathbf{e} = [x_{f2}, x_{p1}, qm_{f2}, qm_{p11}]$	1	0	0	0
$\mathbf{e} = [x_{f2}, x_{p1}, qm_{f2}, qm_{p12}]$	0.9991	0.0007	0	0.0014
$\mathbf{e} = [x_{f2}, x_{p1}, qm_{f2}, qm_{p11}, qm_{p12}]$	**1**	**0**	**0**	**0**
face: 2, fingerprint: 2				
$\mathbf{e} = [x_{f2}, x_{p2}, qm_{f2}]$	0.9819	0.0169	0.0125	0.0213
$\mathbf{e} = [x_{f2}, x_{p2}, qm_{p21}]$	0.9891	0.0107	0.01	0.0114
$\mathbf{e} = [x_{f2}, x_{p2}, qm_{p22}]$	0.9891	0.0102	0.0075	0.0128
$\mathbf{e} = [x_{f2}, x_{p2}, qm_{f2}, qm_{p21}]$	0.9891	0.0107	0.01	0.0114
$\mathbf{e} = [x_{f2}, x_{p2}, qm_{f2}, qm_{p22}]$	0.9891	0.0118	0.015	0.0085
$\mathbf{e} = [x_{f2}, x_{p2}, qm_{f2}, qm_{p21}, qm_{p22}]$	**0.99**	**0.0116**	**0.0175**	**0.0057**
face: 1+2, fingerprint: 1+2				
$\mathbf{e} = [x_{f1}, x_{f2}, x_{p1}, x_{p2}, qm_{f1}, qm_{f2}]$	1	0	0	0
$\mathbf{e} = [x_{f1}, x_{f2}, x_{p1}, x_{p2}, qm_{p11}, qm_{f21}]$	1	0	0	0
$\mathbf{e} = [x_{f1}, x_{f2}, x_{p1}, x_{p2}, qm_{p21}, qm_{f22}]$	1	0	0	0
$\mathbf{e} = [x_{f1}, x_{f2}, x_{p1}, x_{p2}, qm_{f1},$ $, qm_{f2}, qm_{p11}, qm_{f12}]$	**1**	**0**	**0**	**0**
$\mathbf{e} = [x_{f1}, x_{f2}, x_{p1}, x_{p2}, qm_{f1},$ $, qm_{f2}, qm_{p11}, qm_{f12}, qm_{p21}, qm_{p22}]$	1	0	0	0

Table 5.20: Multimodal fusion results for face and fingerprint modalities, development set.

from either face or fingerprint domain, or from both. Listed in Table 5.21, those combinations whose classification results are superior to the baseline presented in Table 5.19, are marked in bold font.

Discussion of the results

The purpose of the fusion experiments was to provide evidential support for the claim that pair-wise dependencies between the components of the evidence vector suffice to observe the benefits of the proposed method of $Q-stack$, and the results presented above confirmed the theoretical expectations worded in Chapter 3. The results clearly show that, as theory anticipates, the use of quality measures in the $Q-stack$ scenario allows for improved classification in respect to the baseline classification

	AC	$HTER$	ER_A	ER_B
face: 1, fingerprint: 1				
$\mathbf{e} = [x_{f1}, x_{p1}, qm_{f1}]$	0.9924	0.0054	0.0021	0.0088
$\mathbf{e} = [x_{f1}, x_{p1}, qm_{p11}]$	**0.999**	**0.0026**	**0.005**	**0.0002**
$\mathbf{e} = [x_{f1}, x_{p1}, qm_{p12}]$	**0.9952**	**0.0037**	**0.0021**	**0.0054**
$\mathbf{e} = [x_{f1}, x_{p1}, qm_{f1}, qm_{p11}]$	0.9946	0.0039	0.0017	0.0062
$\mathbf{e} = [x_{f1}, x_{p1}, qm_{f1}, qm_{p12}]$	0.9943	0.0043	0.0021	0.0064
$\mathbf{e} = [x_{f1}, x_{p1}, qm_{f1}, qm_{p11}, qm_{p12}]$	**0.9989**	**0.0023**	**0.0042**	**0.0004**
face: 1, fingerprint: 2				
$\mathbf{e} = [x_{f1}, x_{p2}, qm_{f1}]$	**0.9917**	**0.0081**	**0.0079**	**0.0083**
$\mathbf{e} = [x_{f1}, x_{p2}, qm_{p21}]$	**0.9931**	**0.0073**	**0.0079**	**0.0067**
$\mathbf{e} = [x_{f1}, x_{p2}, qm_{p22}]$	0.9869	0.0096	0.0042	0.015
$\mathbf{e} = [x_{f1}, x_{p2}, qm_{f1}, qm_{p21}]$	**0.9964**	**0.0069**	**0.0121**	**0.0018**
$\mathbf{e} = [x_{f1}, x_{p2}, qm_{f1}, qm_{p22}]$	0.9866	0.0093	0.0029	0.0157
$\mathbf{e} = [x_{f1}, x_{p2}, qm_{f1}, qm_{p21}, qm_{p22}]$	**0.9974**	**0.0057**	**0.0104**	**0.0009**
face: 2, fingerprint: 1				
$\mathbf{e} = [x_{f2}, x_{p1}, qm_{f2}]$	**0.9967**	**0.0046**	**0.0067**	**0.0026**
$\mathbf{e} = [x_{f2}, x_{p1}, qm_{p11}]$	**0.9982**	**0.0022**	**0.0029**	**0.0015**
$\mathbf{e} = [x_{f2}, x_{p1}, qm_{p12}]$	0.9922	0.0059	0.0029	0.0089
$\mathbf{e} = [x_{f2}, x_{p1}, qm_{f2}, qm_{p11}]$	**0.9983**	**0.0023**	**0.0033**	**0.0013**
$\mathbf{e} = [x_{f2}, x_{p1}, qm_{f2}, qm_{p12}]$	**0.997**	**0.0031**	**0.0033**	**0.003**
$\mathbf{e} = [x_{f2}, x_{p1}, qm_{f2}, qm_{p11}, qm_{p12}]$	**0.9976**	**0.0026**	**0.0029**	**0.0022**
face: 2, fingerprint: 2				
$\mathbf{e} = [x_{f2}, x_{p2}, qm_{f2}]$	0.97	0.023	0.0121	0.0338
$\mathbf{e} = [x_{f2}, x_{p2}, qm_{p21}]$	**0.9934**	**0.0125**	**0.0217**	**0.0034**
$\mathbf{e} = [x_{f2}, x_{p2}, qm_{p22}]$	**0.9923**	**0.0182**	**0.0346**	**0.0019**
$\mathbf{e} = [x_{f2}, x_{p2}, qm_{f2}, qm_{p21}]$	**0.994**	**0.012**	**0.0212**	**0.0028**
$\mathbf{e} = [x_{f2}, x_{p2}, qm_{f2}, qm_{p22}]$	**0.9887**	**0.0126**	**0.0146**	**0.0106**
$\mathbf{e} = [x_{f2}, x_{p2}, qm_{f2}, qm_{p21}, qm_{p22}]$	**0.9947**	**0.0104**	**0.0183**	**0.0025**
face: 1+2, fingerprint: 1+2				
$\mathbf{e} = [x_{f1}, x_{f2}, x_{p1}, x_{p2}, qm_{f1}, qm_{f2}]$	0.9971	0.0019	0.0004	0.0034
$\mathbf{e} = [x_{f1}, x_{f2}, x_{p1}, x_{p2}, qm_{p11}, qm_{f21}]$	**0.9982**	**0.0012**	**0.0004**	**0.0021**
$\mathbf{e} = [x_{f1}, x_{f2}, x_{p1}, x_{p2}, qm_{p21}, qm_{f22}]$	**0.9987**	**0.0011**	**0.0008**	**0.0014**
$\mathbf{e} = [x_{f1}, x_{f2}, x_{p1}, x_{p2}, qm_{f1},]$ $, qm_{f2}, qm_{p11}, qm_{f12}]$	**0.9995**	**0.0008**	**0.0012**	**0.0004**
$\mathbf{e} = [x_{f1}, x_{f2}, x_{p1}, x_{p2}, qm_{f1},]$ $, qm_{f2}, qm_{p11}, qm_{f12}, qm_{p21}, qm_{p22}]$	**0.9988**	**0.001**	**0.0008**	**0.0013**

Table 5.21: Multimodal fusion results for face and fingerprint modalities, evaluation set. The results of classification with quality measures in the proposed $Q - stack$ scheme, superior to the corresponding baseline results, are marked in bold font.

results even when the quality measures are relevant to just one of the scores in the evidence vector. In the results from Table 5.21 this situation is exemplified by all evidence combinations where two scores were used, one coming from the face, and the other from the fingerprint modality.

The results of the fusion experiments also show that even very long evidence vectors still permit achieving very good classification results in the $Q - stack$ scenario. Note that in the experiments reported here we have used evidence vectors counting 8 and 10 elements, and the resulting error rates were better than those of the baseline fusion systems of much smaller dimensionality. It is also worth stressing that in the fusion experiments most of the considered evidence combinations contained the same or larger number of the class-independent quality measures than base classifier scores. For instance, in the evidence vector of $\mathbf{e} = [x_{f1}, x_{f2}, x_{p1}, x_{p2}, qm_{f1}, qm_{f2}, qm_{p11}, qm_{f12}, qm_{p21}, qm_{p22}]$ out of the 10 components only 4 are class-selective. This fact is a direct practical embodiment of the theoretical reasoning about extension of $Q-stack$ to multiple dimensions, given in Chapter 3, Section 3.6. Most notably, the longest evidence vectors containing the most quality measure information resulted in best classification performance, better than the baseline errors listed in Table 5.19.

The results of the multimodal experiments show that $Q - stack$ can systematically improve classification accuracy, even when the used baseline classifier ensembles are already very accurate.

5.5 Conclusions

The experiments reported in this chapter provide a practical example and a pragmatic evaluation of the proposed method of employing class-independent quality information in the classification process, $Q - stack$. We have applied $Q - stack$ to biometric identity verification using single-, multiple-classifier and multimodal systems, involving face image and fingerprint data. In the experiments reported here, classification results using matching similarity scores alone were considered baseline, and were compared to those obtained using the $Q - stack$ framework. We have used a large variety of evidence configurations, and for each of them a properly chosen stacked classifier afforded classification superior to that obtained by the corresponding baseline system. Obtained results show that indeed $Q - stack$ is a powerful and universal method of combining scores and quality measures for improved classification. The improvement in the classification performance in the scheme of $Q - stack$ was obtained exclusively due to the application of quality measures. The quality measures are by themselves are class-independent, individually irrelevant features, and do not offer useful class separation by themselves. In spite of that, in the proposed scheme of $Q - stack$ they are used in much the same way like the class-selective baseline classifications scores. In the context of the baseline classifier scores, quality measures become conditionally relevant class predictors. As such, they are treated like regular classification features for the stacked classifier, with excellent results.

The observed improvements of classification due to the use of quality information are the consequence of the existence of dependencies between the classification scores and quality measures. In the experiments presented in this chapter we have used two estimates of this dependence: Pearson's linear correlation coefficient and pair-wise mutual information. However, it must be noted that neither of these estimates is sufficient to represent the structure of the joint distributions of evidence. While the correlation coefficient conveyed well the characteristic sloping of distributions, its linearity assumption is a limitation for piece-wise correlated data. On the other hand, the mutual information estimate of dependence is oblivious to the type of the dependence present in each individual class. Using any of these estimates alone for prediction of $Q - stack$ performance could lead to mistakes and instead a data-driven approach is postulated. As the presented results show, indeed observed class separation proved to be a good predictor of classification performance on the evaluation data

in the $Q-stack$ scenario.

The experiments also showed that appropriate choice of the stacked classifier in the $Q-stack$ method is crucial for the success of the proposed method. This result is nothing but a consequence of the *no free lunch theorem* [42] which states that the classifier must be chosen according to the classification problem, and no classifier is a priori better than another. In our experiences presented here it was of particular importance that the amount of training data must be sufficient to train all free parameters that define the decision hyper-surface in the evidence space. If the amount of available data is not sufficient, a choice of a stacked classifier of restrictive complexity is more appropriate than running a risk of classifier overtraining. This risk of overtraining increases with the growing number of components in the evidence vector, which automatically increase the dimensionality of the classification space. The effect is well-known in the pattern recognition literature as the *dimensionality curse* [42]. Another conclusion from the practical implementation of $Q-stack$ presented here is that in order to properly train the stacked classifier the structure of the development set should be as close as possible to the structure of the evaluation set, otherwise modeling artifacts may impact negatively the generalization properties of the classifier. Again, this is old news from the viewpoint of pattern classification.

We wish to note here that the purpose of the experiments reported here was to show a systematic improvement in classification thanks to the application of the proposed method of $Q-stack$. Because of the amount and unbalanced nature of available data it is not possible to make claims as to statistical significance of reported differences in classification performance between compared systems. This is for two reasons. First, calculating statistical significance intervals for experiments in biometric identity verification is not an obvious task. Although the available data volume in the evaluation set is fairly large, one must keep in mind that all the samples are not independently drawn from some unknown population. On the contrary, all the available samples come from a much smaller pool of biometric data donors (users). For instance, in the case of the evaluation part of the Biosec database used in our experiments the total number of score samples is 13575 but all these samples originate from only 150 different subjects. Therefore it is not evident how to represent this complicated structure in the statistical significance analysis. Second reason is that for very accurate classifiers and their combinations the resulting errors are due to too small a number of erratic decisions to allow a sound statistical significance analysis. Note that in the experiments involving the fingerprint modality already the baseline classifiers are very accurate. This corresponds to a very limited number of baseline errors. Although applying $Q-stack$ in our experiments led to a systematic and obvious reduction of the number of misclassified samples, their objectively small volume cannot be used as a statistically valid sample for analysis. For instance, the smallest baseline fusion $HTER$ for all baseline scores is 0.16% (Table 5.19). This corresponds to circa 22 erroneous classifier decision in the pool of 13575 classified samples. Using quality measures in the $Q-stack$ scheme halved the number of misclassifications to reach $HTER = 0.08\%$, yet objectively that sums up to a difference of just 11 correct classification decisions more in a pool of data three orders of magnitude larger. It is hardly enough to convince a statistician, yet the reported results show systematic classification improvements on the same database and using the same protocol. This fact shows that $Q-stack$ is indeed a theoretically sound and practically applicable framework of classification with quality measures.

5.6 Summary

In this chapter we have presented an experimental evaluation of the framework of $Q-stack$ applied to the task of biometric identity verification. In our experiments we have used face and fingerprint

data from the Biosec database. For the face modality, we have used two baseline classifiers and we have proposed two algorithms for face image quality estimation. Each of these quality measures was designed to deliver quality estimates that are dependent on the classification scores coming from either of the considered baseline classifiers. For the fingerprint modality, we have applied the publicly available NFIS2 software package to perform minutiae-based matching of fingerprints originating from two different acquisition devices: an optical and a thermal sensor. We have used two fingerprint quality estimation algorithm, one based on spectral analysis of the fingerprint image, and another one based on a classification step. The latter quality measure algorithm, NFIQ, is part of the NFIS2 package.

In our experiments we have considered single-classifier, multiple-classifier, and multimodal combinations of classifiers. The goal of the experiments was to compare the classification results obtained using the baseline systems without the use of quality measures with the results obtained using the $Q-stack$ framework. We have used three different stacked classifiers in the $Q-stack$ architecture, a generative Bayesian classifier, a linear discriminative classifier, and a non-linear discriminative classifier. For each combination of evidence corresponding to the considered single-, multiple-classifier and multimodal architectures the proposed framework of $Q-stack$ allowed for classification at lower error rates than it was the case for the baseline classifiers or their combinations.

We have demonstrated the application of the credence-based method of error prediction, presented in Chapter 4, to the task of face and fingerprint matching. We have shown that credence estimates produced by a stacked Bayesian classifier employed in the $Q-stack$ framework allowed for accurate identification and rejection of potential misclassifications. We have shown that the considered credence estimates are satisfying the evaluation criteria of monotonicity and accountability, presented in Chapter 4. The proposed method of credence-based error prediction was shown to perform well in the presence as well as in the absence of quality measures.

6 Conclusions

In this thesis we have considered the use of quality information for improved classification performance, with application to biometric identity verification. We have noticed the impact of the statistical dependencies between classification features on class separation. We have shown that augmenting the dimensionality of a classification problem with class-independent quality information can be beneficial to class separation, provided that the class-independent quality features sport statistical dependencies on the class-selective features. As a consequence of the existence of these dependencies the quality measures, by themselves irrelevant to the classification process, become conditionally relevant class predictors. We have developed $Q - stack$, a stacking-based approach towards classification with class-independent quality measures, which provides a general unifying framework for quality-based classification in single-, multi-classifier, and multimodal classification systems. We have discussed how alternative approaches towards classification with quality measures can be seen as special cases of $Q - stack$.

We have discussed the topic of estimating probability of error in single classification decisions and we proposed an approach to error prediction based on credence estimation, nested in the subjective Bayesian interpretation of probability. We have shown how a special case of $Q - stack$ that employs a Bayesian stacked classifier can be used to generate credence estimates in single- and multiple classifier architectures, both when quality measures are present and when they are absent. We have explained why the proposed scheme can predict misclassifications more accurately than traditional approach, where only the baseline classifier is used. Finally, we have proposed objective evaluation criteria for credence estimates and we have discussed how the use of credence estimates for rejection of misclassifications together with an appropriate repair strategy can help further reduce error rates to a preset target level.

We have demonstrated the benefits resulting from the application of the proposed framework of $Q - stack$ and of the related error prediction techniques in biometric identity verification system based on the face and fingerprint modalities. Results obtained using real biometric data confirm our theoretical predictions and show that the proposed framework of classification with quality measures systematically outperforms the corresponding baseline systems. We have also shown that the proposed credence-based method of error prediction and rejection allows for accurate prediction

of errors of the considered biometric verification systems and for rejecting unreliable decisions, resulting in systematic gains in classification accuracy.

6.1 $Q-stack$: stacking-based classification with class-independent information

Quality measures quantify the impact of extraneous conditions present during signal acquisition that change the qualitative properties of observed data. As such, by default they do not convey class-related information. In the case of biometric signals, the very same person can be a genuine claimant or an imposter, depending on the identity claims he makes. Yet the identity claims he makes cannot change the quality of the biometric signals collected, and relevant quality measure is therefore class-independent.

The use of class-independent information in pattern classification is not intuitive. Routinely, in pattern classification systems features that grant best possible class separation are preferred and sought. By definition, class-independent quality measures alone provide no class separation and they might seem of limited use for classification - they are individually irrelevant features. In this thesis we notice that for any given set of marginal class-conditional distributions of multiple observed features, the corresponding multivariate class-conditional joint feature distributions depend on the statistical dependencies between the individual features. Consequently, the separation between classes in the multivariate feature space also varies with these dependencies, while the marginal feature distributions remain unchanged. The same phenomenon applies if one or more of the features are class-independent.In this thesis we have shown that the dependency between class-dependent and class-independent classification features offers better class separation than it is the case when these features are statistically independent, given that the dependency is the same for both considered classes. In this sense, the individually irrelevant quality measures become conditionally relevant to the classification process. Further we have discussed the theoretical limitations of the approach, and extended these findings to cases where the dependence is not equal for all classes.

In Chapter 3 we have proposed to combine the classifier similarity scores and the quality measures into evidence vectors, which are then used as feature vectors by a higher-level stacked classifier. The proposed method is inspired by the classifier stacking approach, hence the name $Q - stack$. Given that the dependence between the components of the feature vector is not class-dependent, it guarantees a better class separation than using the baseline classifier scores alone. This improved class separation, given a proper choice of the stacked classifier, translates into higher classification accuracy of the $Q - stack$ ensemble in comparison with the baseline classifier, which does not use quality information. The choice of the stacked classifier is not restricted by any theoretical considerations - it must be chosen according to the observed evidence vectors with respect of the best practices in pattern classification.

We have shown analytically that the presented effect indeed increases the separation between multivariate normal distributions. However, given the high complexity of real non-Gaussian multivariate data, analytical computation of the relations between the feature dependencies and class separation may not be practical, if at all possible. In such applications, we propose a data-driven approach, where the improvements in classification due to the use of the $Q - stack$ scheme in comparison with the baseline classifiers are gauged by actually classifying available development data in the evidence space of similarity scores and quality measures. We have demonstrated the advantages of $Q - stack$ using synthetic and real data sets.

We have demonstrated that the proposed method of $Q - stack$ is a generalization of previously described methods of using quality measures in classification. Existing prior art makes a clear dis-

tinction between the use of quality-measures in single- and multiple-classifier systems. In both cases, quality information was used as a meta-information, distinctly different from the actual classification features. In the prior single-classifier approaches, quality information was used to select from a set of discrete classifiers, models, or decision thresholds used for classification. In the previously proposed multi-classifier scenarios, the use of quality measures was governed by an intuitive notion that out of the given pool of classifiers or modalities, only some are going to be negatively affected by a particular quality degradation. Consequently, quality measures were used to choose (or assign a higher weight to) the classifier or modality that is least likely to be affected. In this work we show that all these methods in fact attempt to create an approximation of the optimal $Q - stack$ decision boundary in the evidence space, using the intuitive notion of quality as data-specific heuristics.

As opposed to the prior alternative approaches, whose heuristic part renders them not generalizable to unseen data or classifier configurations, the proposed method of $Q - stack$ strips quality measures of its meta-informative meaning. Instead, we treat quality measures as regular classification features to the stacked classifier, together with the baseline classifier scores. We demonstrate that the framework of $Q - stack$ can be directly applied to single-, multiple-classifier, and multimodal classification architectures alike. The advantage of the proposed method in comparison to alternative approaches is that the observed gains in classification accuracy have a consistent mathematical explanation, independent on the type and number of baseline classifiers or nature of classified signals used. The theoretical foundations are also not dependent on the context of biometric applications and can therefore be freely applied to any classification task where quality of the signals is of importance.

The main advantages of the proposed framework of classification with quality measures include:

- Proposed method allows for consistent reduction of classification errors thanks to the use of quality measures in single-, multiple-classifier, and multimodal classifier systems.
- The method allows for seamless incorporation of multiple quality measures originating from different modalities into the classification process.
- It is a general classification framework. In the presence of quality measures it encompasses existing approaches to classification with quality information found in the literature. In the absence of quality measures the method turns into a classical stacked generalizer.
- The method can be applied regardless of the nature or modality of the observed signals and independent of the used baseline classifiers.
- The use of quality measures that do not sport a dependency on classification scores does not reduce the class separation. For this reason irrelevant quality measures cannot impair the classification performance, given that an appropriate stacked classifier is applied.

The limitations of the proposed framework include:

- Addition of quality measures to the evidence vector increases the dimensionality of the classification problem faced by the stacked classifier. In such situations the performance of $Q - stack$ is sensitive to a proper choice of the stacked classifier and may be prone to overfitting.
- In theory, if the dependence between quality measures and class-selective features are different for each class, certain configurations of evidence may actually reduce class separation. Such scenarios can be avoided by applying the data-driven approach proposed in this thesis.

Further, the proposed classification framework can take advantage of existing methods of signal-, or feature normalization and conditioning prior to the classification process. As a general rule, any improvement in the performance of the baseline classifier is expected to result in improved performance of the $Q - stack$ ensemble since it improves the class separation of the baseline classifier scores. However, we did not pursue that topic in this thesis. Appropriate normalization and conditioning techniques can also be applied to the actual evidence, depending on the type of the stacked classifier employed. In this thesis we have used simple score and quality measure normalization techniques which aimed at conditioning the features for the use with an SVM classifier. However, more elaborate normalization techniques can be well employed, as discussed in Section 2.

Throughout the thesis we have used a consistent assumption of equal class priors and equal error costs associated with both classes. This simplifying assumption served well in order to keep presented results intuitively accessible, and does not at all interfere with any of the aspects of the proposed classification framework. In terms of classifier design, incorporating uneven class priors and error costs would simply shift the $Q-stack$ decision boundary in the evidence space, identically as it would happen for any classifier that operates in a multivariate feature space.

6.2 Development and proper choice of quality measures

As we have elicited, the advantage of using class-independent quality information in biometric classification emerges from the existence of statistical dependencies between the class-selective similarity scores coming from the baseline classifier or classifiers, and the quality measures. Therefore it is important to bear in mind that rather than developing stand-alone quality measures that measure some objective properties of the observed signals, attention must be paid to the classifier-quality measure ensembles. Only a quality measure that aptly captures the qualitative aspects of the signal, subsequently used by the classifier for class separation, can prove to be of use in reducing classification errors.

In this thesis we have illustrated the process of creating a classifier-quality measure ensemble using the example of face image quality measures. We have also shown how to select an appropriate quality measure when the classification features and baseline classifiers are chosen a priori, using an example of fingerprint-based biometric identity verification.

6.3 Credence-based error prediction and rejection

In many applications of pattern classification the ultimate measure of classification performance is the average error rate. In the process of computing an error rate the total number of wrongly classified decisions is important, but which particular decisions sum up to this number is of little importance. In biometric identity verification it is of paramount significance which of the decisions are correct and which are incorrect since each of the decisions is linked to a particular individual whose identity is in question. It is therefore necessary to not only predict the error rates of a biometric identity verification system but also to predict which individual decisions are likely to be correct: it is necessary to find an estimate of the probability of correctness of a single decision.

A single classification decision is a single probabilistic event, and most probabilistic systems do not include the concept of single event probability. In Chapter 2 we provide a review of most common interpretations of probability, and we show that the interpretation most relevant to this work is Bayesian subjectivism. In the light of the subjective Bayesian interpretation, probability of a single event is a degree of belief, or credence in the occurrence of this event. We therefore propose

to consider single classification decisions as probabilistic events and to use the credence values as the error predictors for single classification decisions.

In Chapter 4 we have presented a credence-based single error prediction scheme which is a natural extension of the $Q - stack$ framework. In the $Q - stack$ architecture, a Bayes stacked classifier produces posterior probabilities of an event that the observed evidence vector comes from a chosen class, given the evidence. This posterior probability is regarded as credence in the correctness of the decision, and can be used to predict and reject decisions that fall below a certain credence threshold. We have shown that such error prediction strategy allows for accurate rejection of potentially erroneous classification decisions from the total decision pool, thus increasing the classification accuracy. The proposed architecture bases on the posterior class probabilities estimated by the stacked classifier, as opposed to the estimates produced by the baseline classifier, which is the case in traditional approaches. The stacked classifier offers better accuracy in assigning credences to decisions than the baseline classifier due to its property of reducing the baseline classifier bias, and due to the inclusion of quality information in the classification process.

Further, we have proposed two criteria to evaluate the obtained credence estimates, a monotonicity criterion and an accountability criterion. The monotonicity criterion demands that the classification error decreases as the credence threshold increases. This criterion corresponds to an intuitive notion that the more credible the classification decisions are required, the higher their classification accuracy must be. The accountability criterion states that the mean credence estimate in the given pool of data must account for the actual accuracy over the considered data set, observed a posteriori. The accountability criterion is a formalization of an intuition that a classifier must produce decisions whose credence estimates will on average reflect and correspond to the observed classification accuracy.

We have shown how the credence-based method of error prediction can be practically deployed to reduce classification error to a target level by applying a suitable repair strategy to all rejected, unreliable decisions. We have discussed the fact that thus obtained benefit of error reduction comes at a cost of repeated data acquisition and classification, and that this cost can be measured by the percentage of decisions that must be discarded from the total decision pool in order to reach the target accuracy. We have shown that both the benefits of attaining a higher classification accuracy and the cost of discarding a necessary volume of unreliable decisions can be estimated using the proposed tools for credence evaluation.

Since the credence-based error prediction is a functional extension of the $Q - stack$ scheme where a Bayesian classifier is applied as a stacked classifier, the method inherits the advantages and limitations of the $Q - stack$ framework. The advantages of the proposed method include:

- Credence estimates can be computed for single-, multiple-classifier, and multimodal classifier systems alike.
- The method can gracefully incorporate (possibly multiple) quality measures.
- The method is capable of producing credence estimates for arbitrary type of baseline classifiers as long as the baseline classifiers return similarity scores.
- The application of the credence-based error prediction scheme is possible regardless of the classified data type or nature.
- Credence estimates can be used for identifying and discarding unreliable decisions prior to applying a repair strategy. In this scenario the classification accuracy can be increased to a desired target level.

The limitations of the proposed method of error prediction include:

- The method requires the use of the estimates of class-dependent joint distributions of evidence. These estimates must be obtained from the available data samples. Although estimation of statistical data distributions from samples is a well-known research area, the choice of the right distribution estimation techniques may be challenging, in particular for evidence vectors containing a large number of components.

As we have shown in this thesis, proposed method of single error prediction allows for detection of potentially unreliable and uncertain classification decisions. Eliminating these decisions from the decision pool allows for increasing the accuracy of classification. In the case of decision rejection an appropriate repair strategy can be applied. For instance, a repeated presentation can be requested, or the system may switch to alternative modalities or classifiers used. Another repair strategy may involve an intervention of a human supervisor. In such case, a probabilistic and intuitively understood by humans credence estimate can be used to help the human supervisor in taking the decision.

6.4 Practical evaluation using a multimodal biometric database

In Chapter 5 we have presented an experimental evaluation of the proposed framework of $Q-stack$ using face and fingerprint data from the Biosec database. Using single-, multiple-classifier and multimodal configurations of two baseline face matchers and two baseline fingerprint matchers we have created a set of baseline systems as a reference point for a comparison with the $Q-stack$ scheme. For the face modality, we have proposed a set of face image quality measures which we have designed to sport dependencies on the chosen baseline face matchers. For the fingerprint modality, we have used a publicly available fingerprint matcher from the NFIS2 software package and we have selected two relevant quality measures, proposed in the prior art. We have elicited the mechanisms of creating a classifier-quality measure ensemble, which we have successfully used in the $Q-stack$ framework. As we have shown in Chapter 5, for all single-, multiple-classifier and multimodal configurations the proposed framework of $Q-stack$ outperformed the corresponding baseline systems in terms of total accuracy and $HTER$. The best performing system overall was a system that involved all available face and fingerprint classifiers with relevant quality measures, which outperformed the corresponding baseline system by an order of error magnitude.

As we have shown in this thesis, practical application of $Q-stack$ is straightforward and follows the very same principles regardless of the number of baseline modalities and classifiers involved. However, care must be taken as the dimensionality of the evidence vectors grows. In this work we have shown $Q-stack$ applied to evidence vectors containing high number of components, most of which were class-independent quality measures. In this situations, some of the stacked classifiers considered in our experiments suffered from overtraining. It can be well anticipated that for an even greater number of involved evidence components a proper estimation of classifier parameters may become hard due to the problem of *dimensionality curse*. In such cases it may be appropriate to reduce the dimensionality of the feature vector or to provide additional observations to augment the training data pool. The reduction in dimensionality of evidence vector may concern only the quality measures, and in such cases compression- or discretization-based approaches can be employed [65, 121].

Finally, we have illustrated the proposed credence-based error prediction method using real biometric face and fingerprint data. We have demonstrated the performance of the presented method, applied to unimodal, multiple-classifier, and multimodal biometric verification architectures. For

each of the architectures, we have used evidence vectors with and without quality measures. In each considered system configuration the obtained credence estimates have been successfully evaluated using the proposed evaluation criteria.

6.5 Future work

In this thesis we have centered our interest around the development of classification methods with quality measures for a dichotomization problem. This focus was motivated by the target application considered in this work, biometric identity verification. Extending the framework of $Q - stack$ to multiple-class problems is a natural continuation of the efforts presented in this dissertation. While certain aspects of presented methodology, for instance the construction of a stacking-based classifier ensemble, are expected to transfer smoothly to multiple-class problems, others may require a more considerable theoretical work, for instance the problem of credence estimation.

In Chapter 3 we have focused on the use of class-independent features on classification, however we have also mentioned that the presented results are relevant to the problem of analyzing the impact of dependencies between class-selective features on class separability. We have also touched on this topic in Appendix A. Although this issue was considered a satellite topic in this thesis, it remains an important and largely unresolved theoretical research subject. It will attract our attention and efforts in the future.

In this thesis we have focused on using the dependencies between signal quality measures and the baseline classification scores for improved classification. While we believe that this is the most straightforward way of mitigating the negative impact of signal noise on classification performance, we envisage that this impact can be addressed also on the level of baseline classification features. Namely, we speculate that quality measures, which would quantify the quality of baseline classifier features, could be directly included in the baseline classification process. Thus constructed baseline classifier could exploit the dependencies between the features and feature-level quality measures in a similar fashion as the $Q - stack$ method proposed in this thesis does at the score level.

Another problem that deserves continued attention is construction of classifier-quality measure ensembles for different signal types, in particular to other biometric modalities. While many fingerprint quality measures exist, the problem of estimating quality of face images in the context of a particular applied classifier is largely unsolved. We hope that the attempts to improve face verification with quality measures we have demonstrated in this dissertation will help to deepen the understanding of what aspects of a face image facilitates its recognition. In this thesis we focused on the tasks of face and fingerprint matching, but we intend to extend the application of the findings of this thesis to other signals, patterns and modalities.

Appendix A

Analysis of KL distance between two bivariate normal distributions

A bivariate conditional distributions of $\mathbf{e} = [x, qm]|A$ and $\mathbf{e} = [x, qm]|B$ whose marginal distributions are normal is given by [82]:

$$\begin{aligned} p(\mathbf{e} = [x, qm]|A) &= \frac{1}{2\pi\sigma_{x,A}\sigma_{qm,A}\sqrt{1-\rho^2}} e^{\left(-\frac{1}{2-2\rho^2}\left(\frac{x^2}{\sigma^2_{x,A}} + \frac{qm^2}{\sigma^2_{qm,A}} - \frac{2x\cdot qm}{\sigma_{x,A}\sigma_{qm,A}}\right)\right)} \\ p(\mathbf{e} = [x, qm]|B) &= \frac{1}{2\pi\sigma_{x,B}\sigma_{qm,B}\sqrt{1-\rho^2}} e^{\left(-\frac{1}{2-2\rho^2}\left(\frac{(x-\mu_{x,B})^2}{\sigma^2_{x,B}} + \frac{(qm-\mu_{qm,B})^2}{\sigma^2_{qm,B}} - \frac{2(x-\mu_{x,B})(qm-\mu_{qm,B})}{\sigma_{x,B}\sigma_{qm,B}}\right)\right)}. \end{aligned} \tag{A.1}$$

In Equation A.1 the correlation coefficient ρ is constrained to satisfy $-1 < \rho < 1$ in order to avoid degenerate cases when $\rho = 1 \Leftrightarrow x = qm$ [42]. Assume $\mu_{x,A} = \mu_{qm,A} = 0$ which without a loss of generality places $p(\mathbf{e}|A)$ at the origin of the coordinate system. For simplicity we further assume $\sigma_{x,A} = \sigma_{x,B} = \sigma_x$ and $\sigma_{qm,A} = \sigma_{qm,B} = \sigma_{qm}$.

Consider the Kullback-Leibler distance between two distributions, $u(\mathbf{x})$ and $v(\mathbf{x})$[42]:

$$D_{KL}(u(\mathbf{x}), v(\mathbf{x})) = \int_{-\infty}^{\infty} v(\mathbf{x}) \ln \frac{v(\mathbf{x})}{u(\mathbf{x})} dx \tag{A.2}$$

For multivariate Gaussian distributions Equation A.2 becomes [84]:

$$D_{KL}(p(\mathbf{e}|A), p(\mathbf{e}|B)) = \ln \frac{|\Sigma_A|}{|\Sigma_B|} + tr\left(\Sigma_A^{-1}\Sigma_B\right) + (\mu_A - \mu_B)^T \Sigma_B^{-1} (\mu_A - \mu_B), \tag{A.3}$$

where Σ_A and Σ_B are covariance matrices of A and B and $\mu_A = [mu_{x,A} mu_{qm,A}]^T$, $\mu_B = [mu_{x,B} mu_{qm,B}]^T$. In general case the Kullback-Leibler distance is asymmetric, $D_{KL}(A, B) \neq D_{KL}(B, A)$. In our case $\Sigma_A = \Sigma_B = \Sigma$. In this situation Equation A.4 becomes:

$$D_{KL}(p(\mathbf{e}|A), p(\mathbf{e}|B)) = D_{KL}(p(\mathbf{e}|B), p(\mathbf{e}|A)) = (\mu_B)^T \Sigma^{-1} (\mu_A) + \beta, \tag{A.4}$$

where

$$\beta = \ln \frac{|\Sigma|}{|\Sigma|} + tr\left(\Sigma^{-1}\Sigma\right) = 0 + 2 = 2 = const.$$

The metric therefore becomes symmetric and proportional to divergence between $\mathbf{e}|A$ and $\mathbf{e}|B$. We proceed to show that it has one and only one minimum for $-1 < rho < 1$. First let us transform A.4 into an explicit function of ρ:

$$D_{KL}(\rho) = (\mu_B)^T \Sigma^{-1} (\mu_B) = \frac{1}{1-\rho^2}\left(\frac{\mu_x^2}{\sigma_x^2} + \frac{\mu_{qm}^2}{\sigma_{qm}^2} - \frac{2\rho\mu_x\mu_{qm}}{\sigma_x\sigma_{qm}}\right)$$

In order to find characteristic points of $D_{KL}(\rho)$ compute 1st and 2nd derivative of $D_{KL}(\rho)$:

$$\frac{d}{d\rho}D_{KL}(\rho) = \frac{\rho}{(1-\rho^2)^2}\left(\frac{2\mu_x^2}{\sigma_x^2} + \frac{2\mu_{qm}^2}{\sigma_{qm}^2} - \frac{4\rho\mu_x\mu_{qm}}{\sigma_x\sigma_{qm}}\right) - \frac{2\mu_x\mu_{qm}}{\sigma_x\sigma_{qm}(1-\rho^2)} \tag{A.5}$$

$$\frac{d^2}{d\rho^2}D_{KL}(\rho) = \left(\frac{2\mu_x^2}{\sigma_x^2} + \frac{2\mu_{qm}^2}{\sigma_{qm}^2} - \frac{4\rho\mu_x\mu_{qm}}{\sigma_x\sigma_{qm}}\right)\left(\frac{1}{(1-\rho^2)^2} + \frac{4\rho^2}{(1-\rho^2)^2}\right) - \frac{8\rho\mu_x\mu_{qm}}{\sigma_x\sigma_{qm}(1-\rho^2)} \tag{A.6}$$

In order to find the extrema of $D_{KL}(\rho)$ we need to find ρ for which the first derivative of $D_{KL}(\rho)$ given by A.5 is equal zero.

$$\frac{d}{d\rho}D_{KL}(\rho) = 0 \Rightarrow \frac{1}{(1-\rho^2)}\left(\frac{\rho}{(1-\rho^2)}\left(\frac{2\mu_x^2}{\sigma_x^2} + \frac{2\mu_{qm}^2}{\sigma_{qm}^2} - \frac{4\rho\mu_x\mu_{qm}}{\sigma_x\sigma_{qm}}\right) - \frac{2\mu_x\mu_{qm}}{\sigma_x\sigma_{qm}}\right) = 0 \Rightarrow$$

$$\Rightarrow \frac{\rho}{(1-\rho^2)}\left(\frac{\mu_x^2}{\sigma_x^2} + \frac{\mu_{qm}^2}{\sigma_{qm}^2} - \frac{2\rho\mu_x\mu_{qm}}{\sigma_x\sigma_{qm}}\right) - \frac{\mu_x\mu_{qm}}{\sigma_x\sigma_{qm}} = 0 \Rightarrow$$

$$\Rightarrow (\mu_x\sigma_{qm} - \rho\mu_{qm}\sigma_x)(\mu_{qm}\sigma_x - \rho\mu_x\sigma_{qm}) = 0$$

Solution of the above equation yields:

$$\rho_1 = \frac{\mu_{qm}\sigma_x}{\mu_x\sigma_{qm}}, \rho_2 = \frac{\mu_x\sigma_{qm}}{\mu_{qm}\sigma_x} \tag{A.7}$$

Note that by definition

$$|\rho_1| < 1 \Leftrightarrow |\rho_1| > 1$$

Therefore Equation A.4 has only one stationary point in the range of $-1 > \rho > 1$. Let us assume that $|\rho_1| < 1$. At this point Equation A.6 evaluates to:

$$D''_{KL}(\rho_1) = \frac{2\sigma_{qm}^2\mu_x^4}{(\sigma_{qm}^2\mu_x^2 - \sigma_x^2\mu_{qm}^2)\sigma_x^2}$$

Since $-1 > \rho_1 > 1$ then necessarily

$$\sigma_{qm}^2\mu_x^2 > \sigma_x^2\mu_{qm}^2 \Rightarrow \sigma_{qm}^2\mu_x^2 - \sigma_x^2\mu_{qm}^2 > 0 \Rightarrow D''_{KL}(\rho_1) > 0$$

which indicates that $D_KL(\rho)$ has a minimum at ρ_1. Minimal divergence between distributions indicates that for ρ_1 the information for discrimination between classes A and B is also minimized.

Should $|\rho_2| < 1$ be assumed then ρ_2 would be the one and only one valid solution that minimizes D_{KL}.

Note that this result is identical with that obtained from the error function analysis presented in Section 3.4.1.

Numerical analysis of generic cases: $\mu_{qm,B} \neq 0$, $\Sigma_{qm,A} \neq \Sigma_{qm,B}$

We have given an analytical proof for a specific case when $\Sigma_A = \Sigma_B$. Of course this case is a special one. For $\Sigma_A \ll \Sigma_B$ and $\Sigma_A \gg \Sigma_B$ the divergence between two Gaussian distributions may have more extrema for correlation coefficient $\rho \in (-1, ..., 1)$. However, this proof is of limited relevance to the main topic of this thesis since then the marginal class-conditional distributions in dimension qm would actually not be class-independent. Considering its notational size and complexity we prefer to skip the analytical analysis of this case. Instead, we chose to show a series of numerically evaluated examples together with their graphical representations, which in our opinion is more illustrative and gives the reader a better intuitive understanding of the implications of correlation between classification features.

Case 1: Changing $\mu_{qm,B}$. Figure A.1 presents how Bayes error for joint distributions $p(\mathbf{e}|A)$ and $p(\mathbf{e}|B)$ changes as a function of correlation coefficient ρ. Subsequent curves are drawn for changing mean of the marginal distribution of $p(qm|B)$. Distributions parameters: $\mu_{x,B} = 3$,$\sigma_{x,A} = \sigma_{x,B} = 1$,$\sigma_{qm,A} = \sigma_{qm,B} = 3$, $\mu_{qm,B}$ changing from -5 to 5.

The solid black line in Figure A.1 represents the curve for $\mu_{qm,B} = 0$, when marginals in qm dimension are equal, $p(qm|B)$. Dashed red curves are drawn for $\mu_{qm,B} < 0$ and solid blue curves for $\mu_{qm,B} > 0$. Note that there is still one maximum of Bayes error but for $\mu_{qm,B} \neq 0$ it no longer happens for $\rho = 0$. In respect to the case when $\rho = 0$ the class separation may therefore increase or decrease, depending on the sign of ρ.

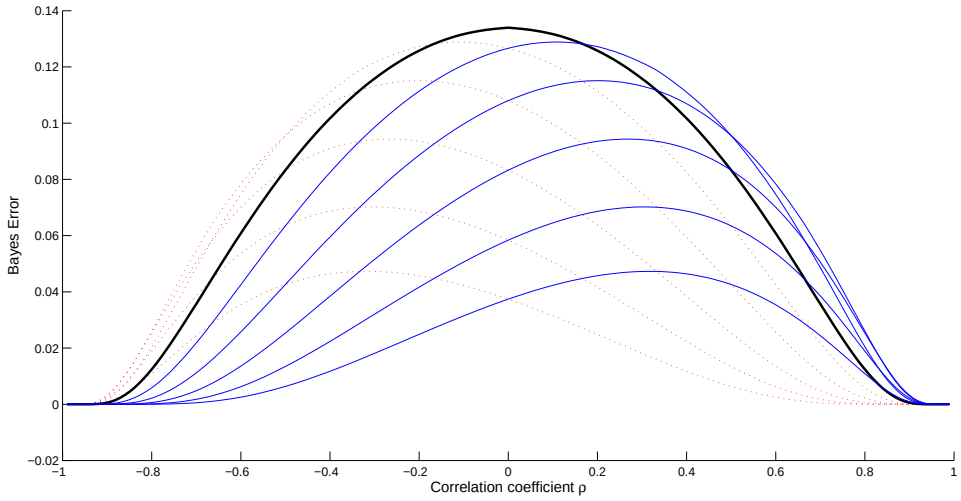


Figure A.1: Effect of correlation ρ, $\mu_{qm,B}$ changing from -5 to 5

Case 2: Changing $\sigma_{qm,B}$. Figure A.2 presents how Bayes error for joint distributions $p(\mathbf{e}|A)$ and $p(\mathbf{e}|B)$ changes as a function of correlation coefficient ρ, when for changing the variance $\sigma^2_{qm,B}$ of the marginal distribution of $p(qm|B)$. Distributions parameters: $\mu_{x,B} = 3$, $\mu_{qm,B} = 3$, $\sigma_{x,A} = \sigma_{x,B} = 1$, $\sigma_{qm,A} = 3$, $\sigma_{qm,B}$ changing from 0.5 to 5.5.

When changing only the variance $\sigma^2_{qm,B}$ of the marginal distribution of $p(qm|B)$, the Bayes error may take more than one extremum in the permitted range of ρ. This corresponds to the case when there is a large difference in variances $\sigma_{qm,A}$ and $\sigma_{qm,B}$. In the example shown in Figure A.2 the case when $\sigma_{qm,A} \gg \sigma_{qm,B}$ is shown with a bold black line. As $\sigma_{qm,B}$ increases to approach $\sigma_{qm,A}$ (bold green curve) the sharp maximum observed for large values of ρ, and the minimum leave the

permitted range of $-1 < \rho < 1$ and the curve sports only one maximum.

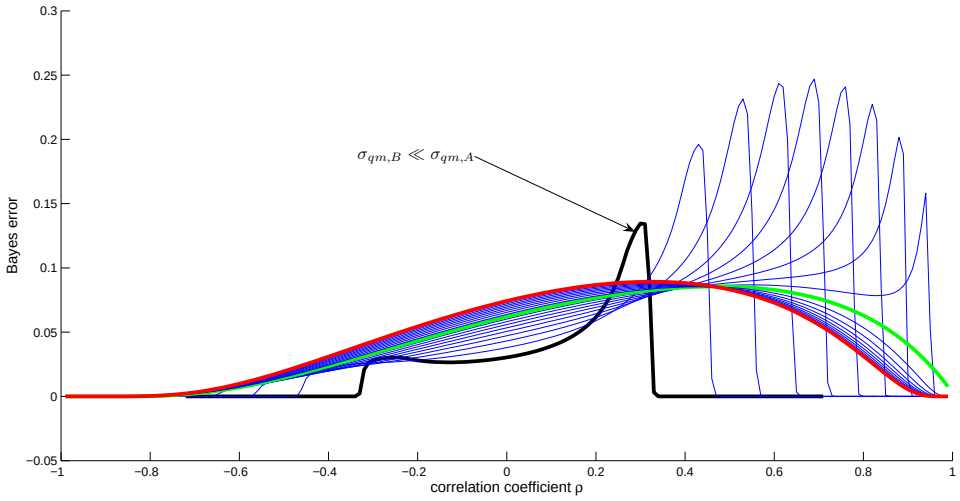


Figure A.2: Effect of correlation ρ, $\sigma_{qm,B}$ changing from -0.1 to 10

As we have demonstrated in Figures A.1 and A.2 the dependency of the class separation on the correlation coefficient ρ is complex and strongly dependent on the actual parameters of the joint distributions $p(\mathbf{e}|A)$ and $p(\mathbf{e}|B)$. This explains the incomplete, conflicting or not fully conclusive reports on the impact of feature correlation on classification performance found in the literature [53, 84, 119]. The complexity of the problem forced us to present also only a fragmentary view on the topic. The issue deserves a more indepth mathematical analysis, but since this thesis focuses on the case when $p(qm|A) = p(qm|B)$ it is of limited relevance here.

Appendix B

$Q - stack$ for multiplicative noise - example

In Chapter 3, Section 3.8 we have demonstrated the principles of $Q - stack$ on the example of normally distributed data, contaminated with additive Gaussian noise. We have suggested that the assumption and knowledge of the noise nature was not important for the proposed method of classification with quality information, $Q - stack$. In this Appendix we will substantiate this claim by showing another example, where the noise is multiplicative rather than additive.

Assume classes A and B that generate x' according to:

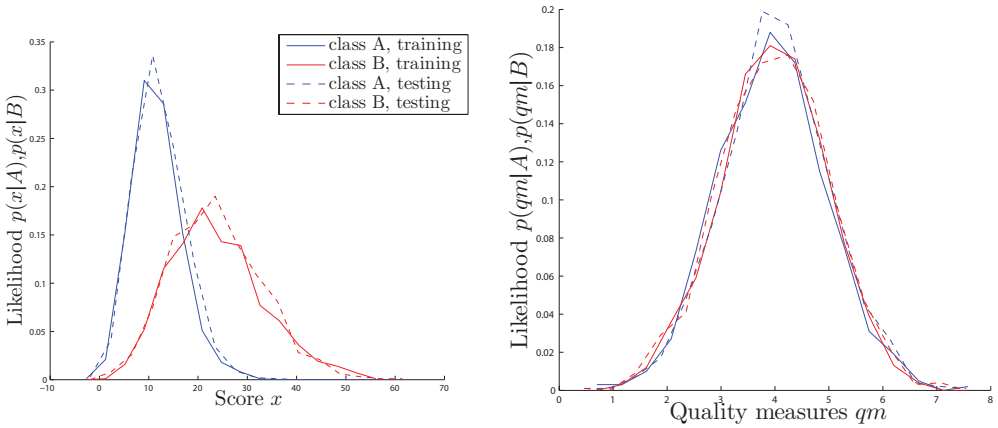
$$\begin{aligned} p(x'|A) &= \mathcal{N}(\mu_{x',A}, \sigma^2_{x',A}), \mu_{x',A} = 3, \sigma^2_{x',A} = 1, \\ p(x'|B) &= \mathcal{N}(\mu_{x',B}, \sigma^2_{x',B}), \mu_{x',B} = 6, \sigma^2_{x',B} = 3 \end{aligned} \tag{B.1}$$

next, assume a process N that generates noise instances n according to

$$p(n) = \mathcal{N}(\mu_N, \sigma^2_N), \mu_N = 4, \sigma^2_N = 1, \tag{B.2}$$

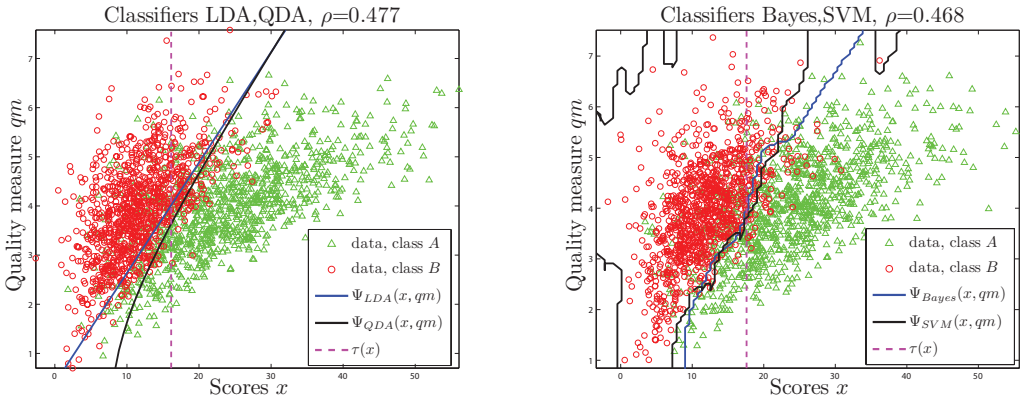
The noise instances n are affecting x' according to the function $x = \Phi(x', n) = n \cdot x'$, generating noisy observations (scores) x. Quality measures qm are collected by measuring the noise directly. For $qm = n$ the marginal class-conditional distributions of x and qm are shown in Figures B.1. The quality measures and classification scores are combined into evidence vectors $\mathbf{e} = [x, qm]$ and classified using four different stacked classifiers (LDA, QDA, $Bayes$, SVM), according to the framework of $Q - stack$ discussed in Chapter 3. Classification in the evidence space using these classifiers is shown graphically in Figure B.2.

Recall from Section 3.8 that we have analyzed the impact of uncertainty in the estimation of the correlation coefficient ρ on class separation, and consequently on classification performance of a $Q - stack$ ensemble, for the case of additive noise. In this Appendix we are presenting a simialr analysis for the case of multiplicative noise. For the details on the experiment design the reader is referred to Section 3.8. The graphs that demonstrate the link between the certainty in estimating quality measures qm are given in Figure B.3. The classification performance by each of the stacked classifiers is expressed in terms of $HTER$ in Table B.1.



(a) Observed scores x (b) Quality measures qm

Figure B.1: Subplot a.:Class-conditional distributions $p(x|A)$ and $p(x|B)$. Subplot b.:Class-conditional distributions $p(qm|A)$ and $p(qm|B)$



(a) Classifiers LDA,QDA (b) Classifiers $Bayes$,SVM

Figure B.2: Classification in the evidence space $\mathbf{e} = [x, qm]$ using (a) LDA, QDA , and (b) $Bayes$, SVM stacked classifiers.

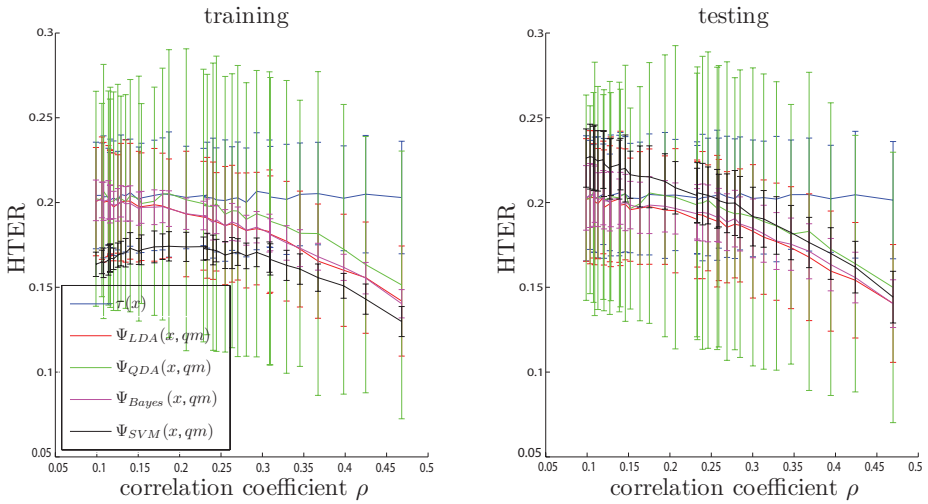


Figure B.3: Impact of the correlation ρ between the observed scores x and the observed quality measures qm , for multiplicative noise.

σ_d^2	0.000	0.400	1.000	2.600	7.000	15.000	20.000
$\overline{\rho(n, qm)}$	1.000	0.843	0.707	0.526	0.353	0.252	0.220
$\overline{\rho(x, qm)}$	0.000	0.007	0.012	0.014	0.019	0.022	0.022
$HTER$							
$\mu_{\tau(qm)}$	0.500	0.502	0.498	0.498	0.501	0.503	0.501
$\sigma_{\tau(qm)}$	0.036	0.036	0.041	0.040	0.039	0.032	0.035
$\mu_{\tau(x)}$	0.201	0.202	0.202	0.203	0.202	0.205	0.203
$\sigma_{\tau(x)}$	0.035	0.032	0.037	0.035	0.033	0.030	0.037
μ_{LDA}	0.141	0.160	0.177	0.191	0.197	0.203	0.202
σ_{LDA}	0.035	0.035	0.035	0.031	0.033	0.033	0.036
μ_{QDA}	0.150	0.173	0.186	0.201	0.199	0.203	0.203
σ_{QDA}	0.080	0.086	0.085	0.087	0.069	0.060	0.061
μ_{Bayes}	0.140	0.163	0.177	0.194	0.197	0.203	0.203
σ_{Bayes}	0.014	0.015	0.017	0.017	0.015	0.016	0.019
μ_{SVM}	0.144	0.170	0.186	0.204	0.215	0.225	0.226
σ_{SVM}	0.015	0.015	0.017	0.017	0.018	0.017	0.017

Table B.1: Selected $HTER$ results from Figure B.3(b), 1000 data points, mean values and standard deviations after 50 repetitions for each value of σ_d^2.

Appendix C

DET curves for the evaluation experiments

In this Appendix we provide the reader with selected DET curves drawn for the $Q-stack$ experiments described in Chapter 5. The DET curves are not a realistic evaluation of the performance of a pattern classification system since they consider only the scores obtained on the evaluation set, disregarding the last step classification step - making actual classification decisions. Yet, it shows what the classification performance of a system *would be* if a particular decision threshold were selected. An advantage of representing the classifier's performance using a DET curves is the accessibility of its visual representation.

The DET curves are organized by the considered baseline classification system:

Face modality, DCT baseline classifier: Figure C.1, (a) $Bayes$, (b) $SVM-lin$, (c) $SVM-rbf$ stacked classifiers.

Face modality, PCA baseline classifier: Figure C.2, (a) $Bayes$, (b) $SVM-lin$, (c) $SVM-rbf$ stacked classifiers.

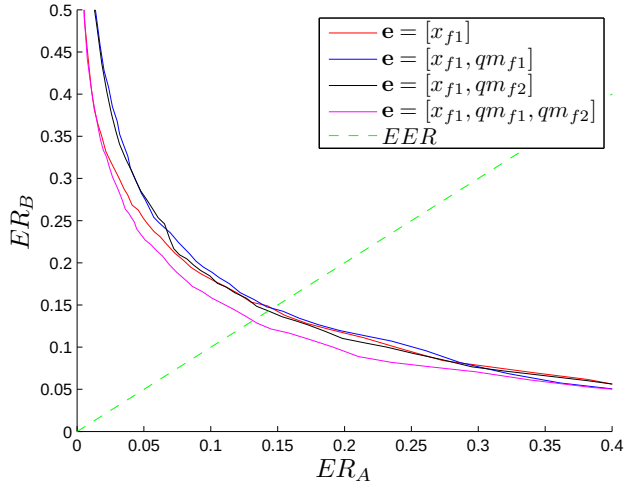
Face modality, fusion of $DCT+PCA$ baseline classifiers: Figure C.3, (a) $Bayes$, (b) $SVM-lin$, (c) $SVM-rbf$ stacked classifiers.

Fingerprint modality, optical: Figure C.4, (a) $Bayes$, (b) $SVM-lin$, (c) $SVM-rbf$ stacked classifiers.

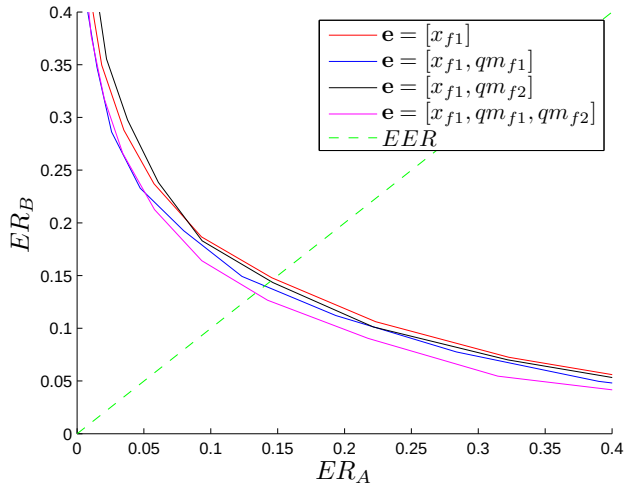
Fingerprint modality, optical: Figure C.5, (a) $Bayes$, (b) $SVM-lin$, (c) $SVM-rbf$ stacked classifiers.

Fingerprint, fusion of optical+optical modalities: Figure C.6, (a) $Bayes$, (b) $SVM-lin$, (c) $SVM-rbf$ stacked classifiers.

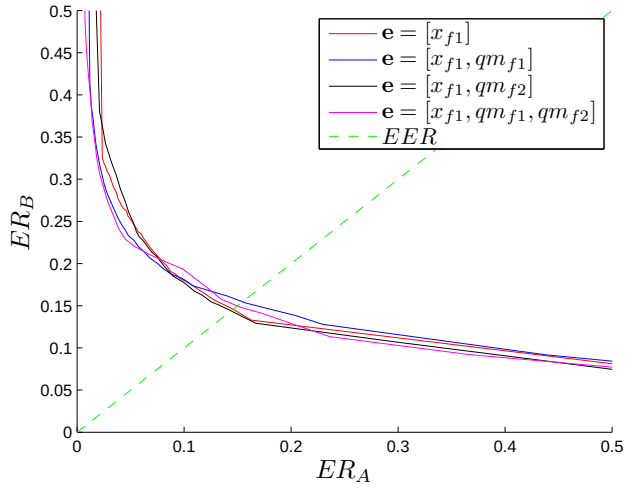
Fingerprint+face, multimodal fusion: Figures C.7, C.8, C.9, C.10, C.11, $SVM-lin$ stacked classifiers.



(a) $x_{f1}, Bayes$

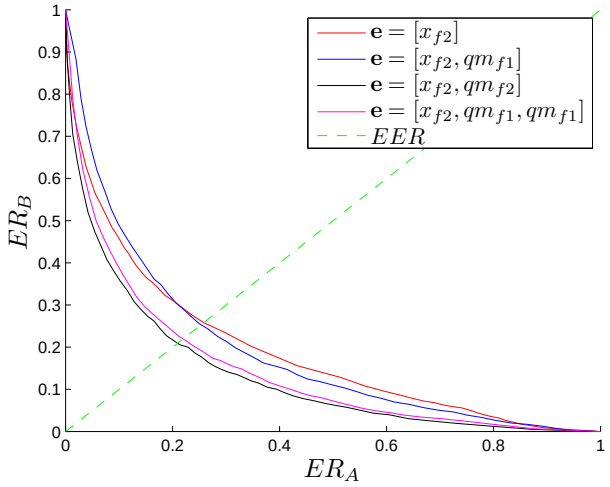


(b) $x_{f1}, SVM - lin$

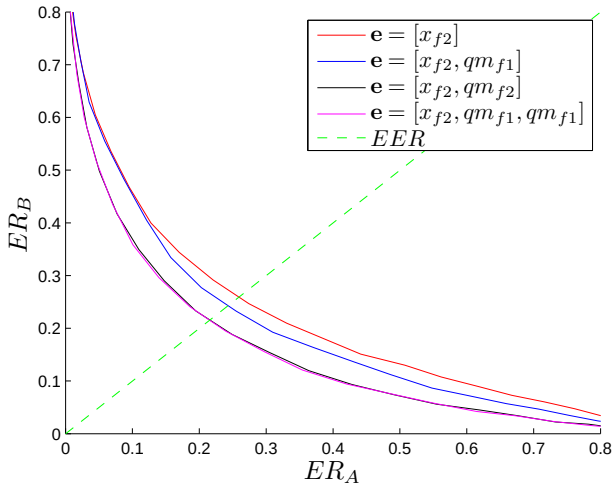


(c) $x_{f1}, SVM - rbf$

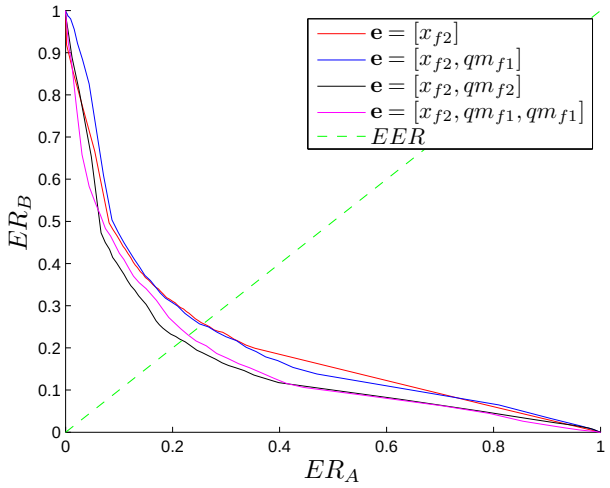
Figure C.1: DET curves for face modality, *DCT* baseline classifier.



(a) $x_{f2}, Bayes$

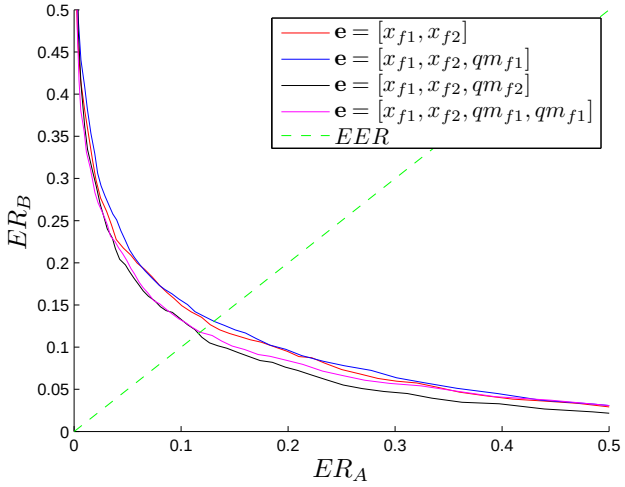


(b) $x_{f2}, SVM - lin$

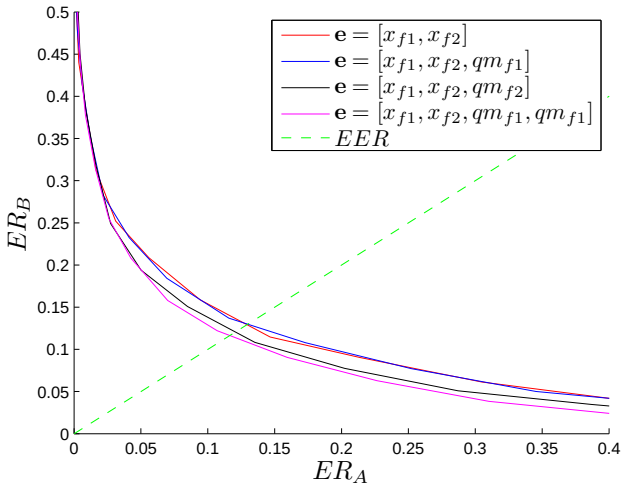


(c) $x_{f2}, SVM - rbf$

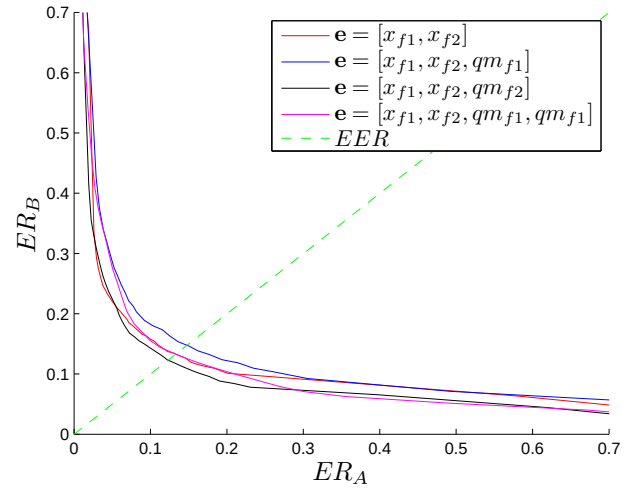
Figure C.2: DET curves for face modality, PCA baseline classifier.



(a) $x_{f1}, x_{f2}, Bayes$

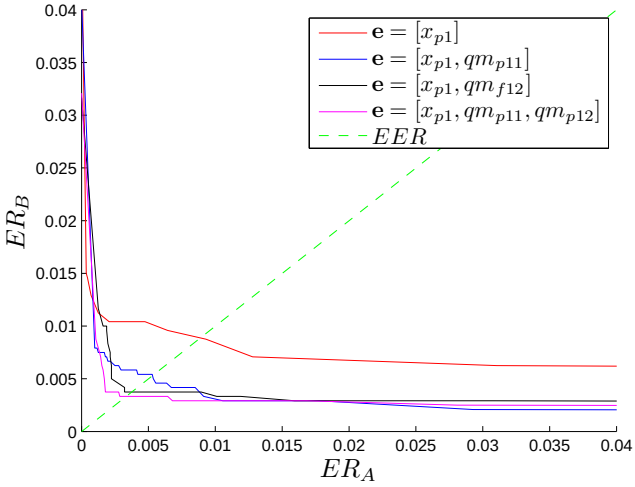


(b) $x_{f1}, x_{f2}, SVM - lin$

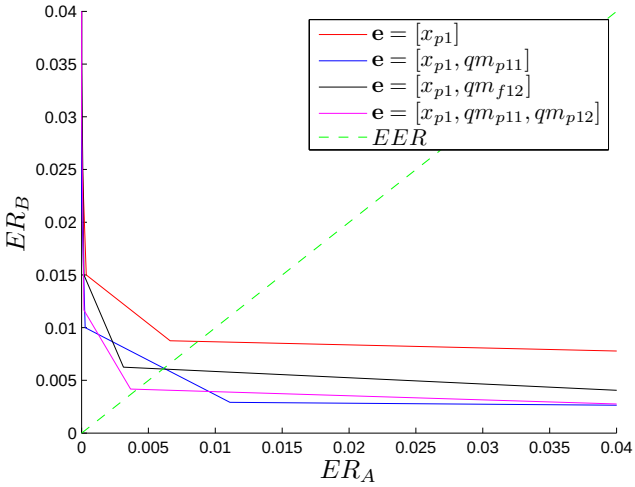


(c) $x_{f1}, x_{f2}, SVM - rbf$

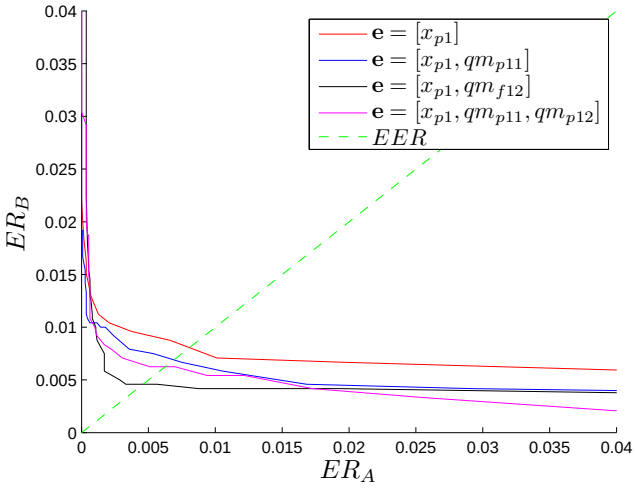
Figure C.3: DET curves for face modality, multiple-classifier face verification involving DCT and PCA baseline classifiers.



(a) $x_{p1}, Bayes$

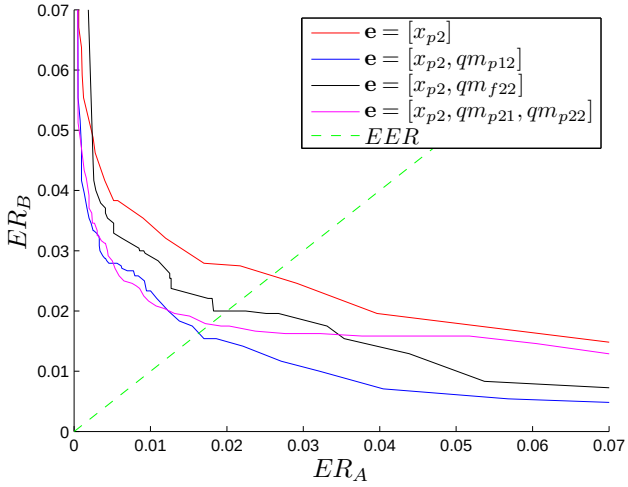


(b) $x_{p1}, SVM - lin$

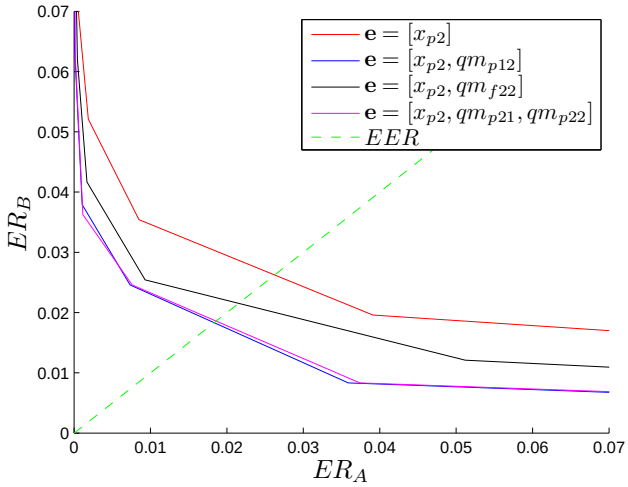


(c) $x_{p1}, SVM - rbf$

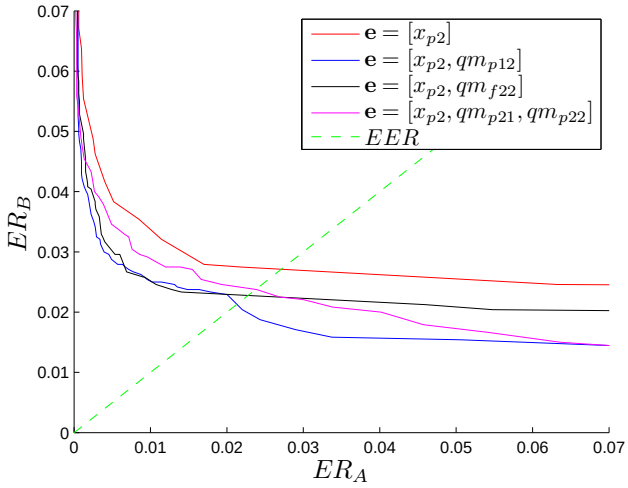
Figure C.4: DET curves for fingerprint modality, optical sensor.



(a) $x_{p2}, Bayes$

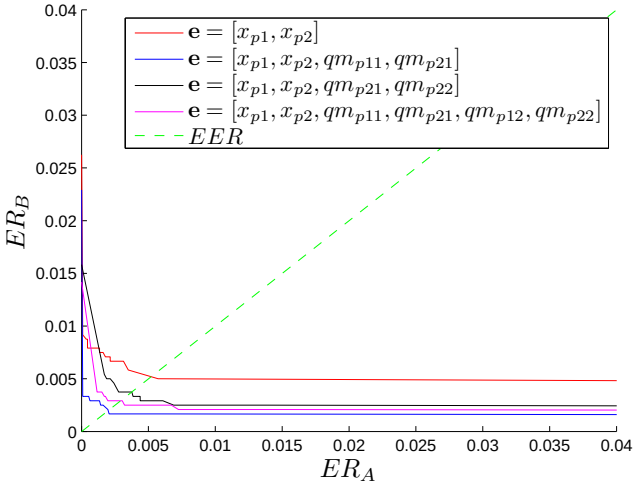


(b) $x_{p2}, SVM - lin$

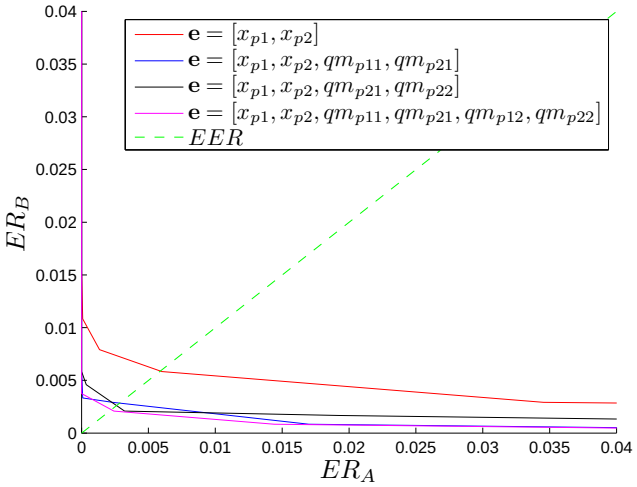


(c) $x_{p2}, SVM - rbf$

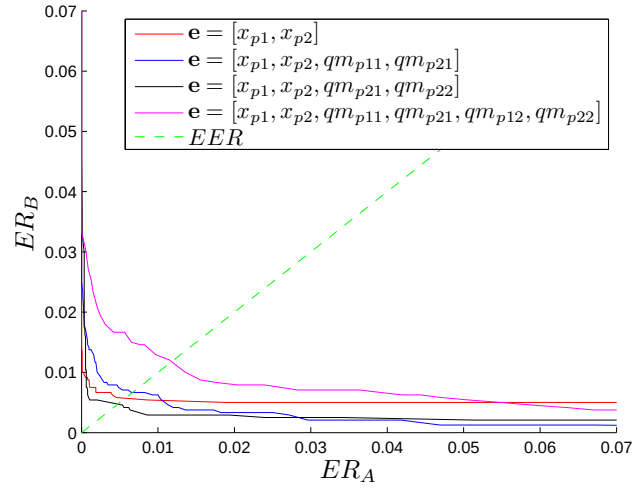
Figure C.5: DET curves for fingerprint modality, thermal sensor.



(a) $x_{p1}, x_{p2}, Bayes$



(b) $x_{p1}, x_{p2}, SVM - lin$



(c) $x_{p1}, x_{p2}, SVM - rbf$

Figure C.6: DET curves for fingerprint multimodal fusion, optical + thermal sensors.

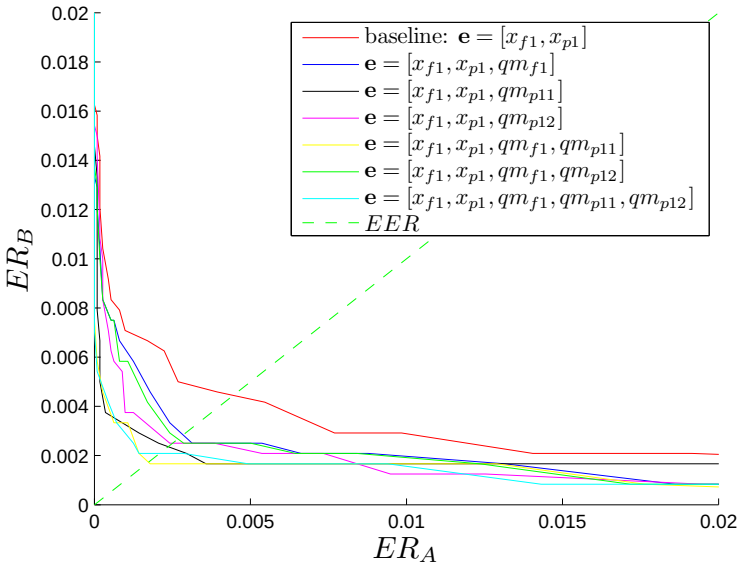


Figure C.7: DET curves for fusion of face and fingerprint modality, $x_{f1}, x_{p1}, SVM - lin$.

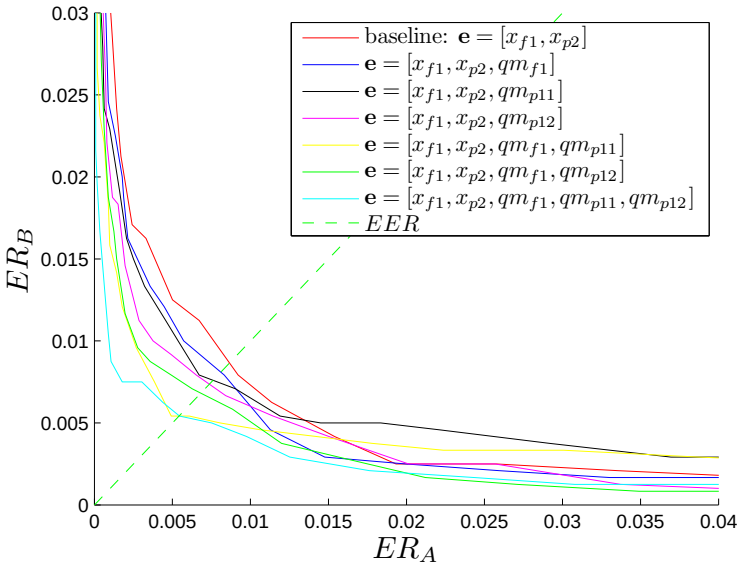


Figure C.8: DET curves for fusion of face and fingerprint modality, $x_{f1}, x_{p2}, SVM - lin$.

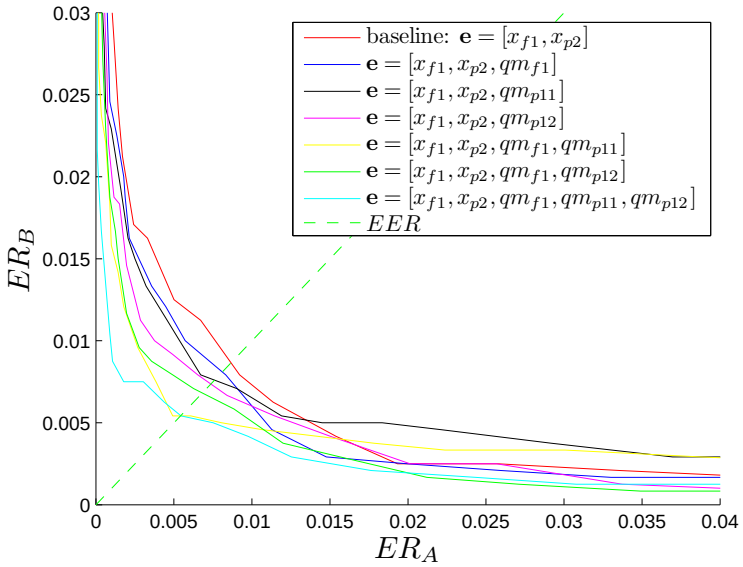


Figure C.9: DET curves for fusion of face and fingerprint modality, $x_{f2}, x_{p1}, SVM - lin$.

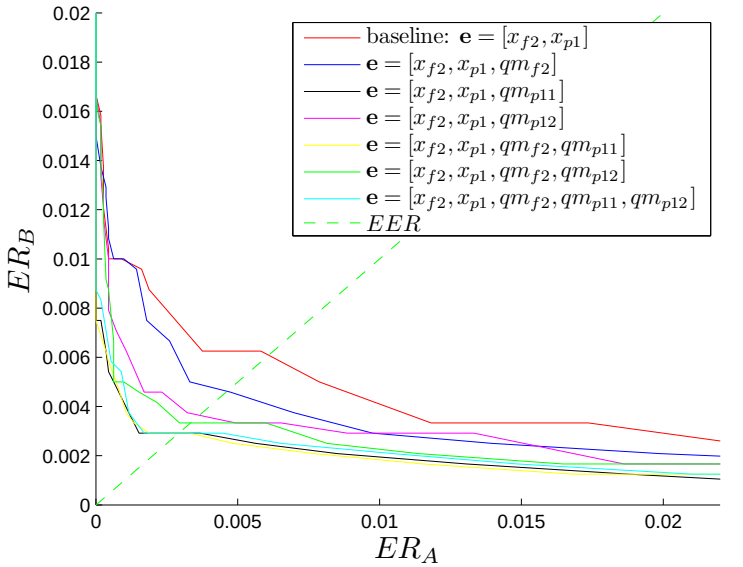
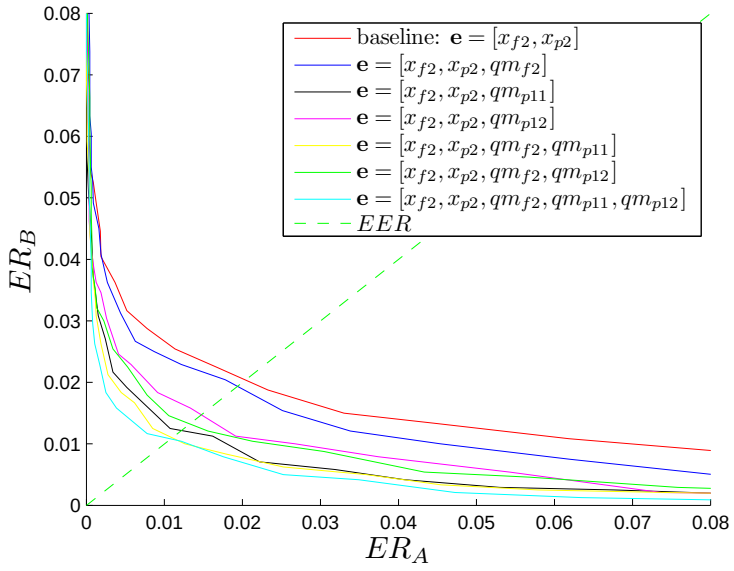


Figure C.10: DET curves for fusion of face and fingerprint modality, $x_{f2}, x_{p2}, SVM - lin$.



(a)

Figure C.11: DET curves for fusion of face and fingerprint modality, $x_{f1}, x_{p1}, x_{f1}, x_{p1}, SVM - lin.$

Bibliography

[1] Technical specifications, ATMEL FCDEMO4. http://www.atmel.com/dyn/products/product_card.asp?part_id=3322.

[2] Biometrics deployment of machine readable travel documents, Annex A: Photograph Guidelines. Technical report, ICAO, May 2004.

[3] Technical specifications, Biometrika FX 2000 fingerprint scanner. http://www.biometrika.it/eng/fx2000.html, 2004.

[4] Biometric data interchange formats - part 5: Face image data, ISO/IEC 19794-5:2005. http://webstore.ansi.org/ansidocstore/product.asp?sku=ISO/IEC+19794-52005.

[5] BioSec Technical Report IST-2002-001766 D5, extended multimodal database and testing protocol, 5. performance evaluation on the multimodal biometric database. Technical report, BIOSEC, 2005.

[6] A. Adler and T. Dembinsky. Human vs. automatic measurement of biometric sample quality. In *Canadian Conference on Computer and Electrical Engineering (CCECE)*, Ottawa, Canada, May 2006.

[7] A. Alexander. *Forensic automatic speaker recognition using Bayesian interpretation and statistical compensation for mismatched conditions*. PhD thesis, Swiss Federal Institute of Technology Lausanne (EPFL), 2005.

[8] F. Alonso-Fernandez, J. Fierrez-Aguilar, and J. Ortega-Garcia. A review of schemes for fingerprint image quality computation. In *Proceedings of COST 275 Workshop - Biometrics based recognition of people over the internet*, Hatfield, UK, 2005.

[9] I. Avcibas, B. Sankur, and K. Sayood. Statistical evaluation of image quality measures. *Journal of Electronic Imaging*, 11(2):206–223, April 2002.

[10] E. Bailly-Bailliére, S. Bengio, F. Bimbot, M. Hamouz, J. Kittler, J. Mariéthoz, J. Matas, K. Messer, V. Popovici, F. Porée, B. Ruiz, and J.-P. Thiran. The BANCA database and evaluation protocol. In J. Kittler and M.S. Nixon, editors, *Proceedings of 4th Int. Conf. on Audio- and Video-Based Biometric Person Authentication (AVBPA)*, volume LNCS 2688, pages 625–638, 2003.

[11] C. Barras and J-L. Gauvain. Feature and score normalization for speaker verification of cellular data. In *International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2003)*, Hong Kong, China, 2003.

[12] A. M. Bazen and R. N. J. Veldhuis. Likelihood-ratio-based biometric verification. *IEEE Transactions on Circuits and Systems for Video Technology*, 14(1), January 2004.

[13] P. N. Belhumeur, J. Hespanha, and D. J. Kriegman. Eigenfaces vs. fisherfaces: Recognition using class specific linear projection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(7):711–720, 1997.

[14] R. E. Bellman. *Dynamic Programming.* Princeton University Press, Princeton, NJ, 1957.

[15] A. P. Benavent, F. E. Ruiz, and J. S. Martinez. EBEM: an entropy-based EM algorithm for Gaussian Mixture Models. In *Proc. of the 18th Intl. Conference on Pattern Recognition*, Hong Kong, August 2006.

[16] S. Bengio, C. Marcel, S. Marcel, and J. Mariethoz. Confidence measures for multimodal identity verification. *Information Fusion*, 3(4):267–276, 2002.

[17] S. Bengio and J. Matriethoz. Biometric person authentication is a multiple classifier problem. In *Proceedings of the 7th International Workshop on Multiple Classifier Systems, MCS'07*, Prague, Czech Republic, May 2007.

[18] E. S. Bigun. Risk analysis of catastrophes using experts' judgements: An empirical study on risk analysis of major civil aircraft accidents in Europe. *European Journal of Operational Research*, 87(3):599–612, December 1995.

[19] E. S. Bigun, J. Bigun, B. Duc, and S. Fischer. Expert conciliation for multi modal person authentication systems by bayesian statistics. In *Proceedings of the First International Conference on Audio- and Video-Based Biometric Person Authentication AVBPA '97*, pages 291–300, London, UK, 1997. Springer-Verlag.

[20] J. Bigün, J. Fiérrez-Aguilar, J.Ortega-Garcia, and J. Gonzalez-Rodriguez. Combining biometric evidence for person authentication. In *Advanced Studies in Biometrics*, pages 1–18, 2003.

[21] J. Bigun, J. Fierrez-Aguilar, J. Ortega-Garcia, and J. Gonzalez-Rodriguez. Multimodal biometric authentication using quality signals in mobile communications. In *Proc. 12th International Conference on Image Analysis and Processing*, Mantova, Italy, 2003.

[22] R. Bose. *Information Theory, Coding and Cryptography.* McGraw Hill, 2002.

[23] L. Breiman. Bagging predictors. *Machine Learning*, 24(2):123–140, 1996.

[24] L. Bruzzone, F. Roli, and S. B. Serpico. An extension of the Jeffreys-Matusita distance to multielass cases for feature selection. *IEEE Transactions on Geoscience and Remote Sensing*, 33(6):1318–1321, November 1995.

[25] W. Campbell, D. Reynolds, J. Campbell, and K. Brady. Estimating and evaluating confidence for forensic speaker recognition. In *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 1, pages 717–720, 2005.

[26] R. Carnap. *Logical Foundations of Probability.* University of Chicago Press, 1950.

[27] O. Yuridia Gonzalez Castillo. Report: Survey about facial image quality. Technical report, Fraunhofer Institute for Computer Graphics Research, Darmstadt, Germany, December 2006.

[28] V. Chatzis, A. G. Bors, and I. Pitas. Multimodal decision-level fusion for person authentication. *IEEE Trans. on Systems, Man, and Cybernetics, Part A: Systems and Humans*, 29(6):674–680, 1999.

[29] Y. Chen, S. Dass, and A. Jain. Fingerprint quality indices for predicting authentication performance. In *Proc. of Audio- and Video-based Biometric Person Authentication (AVBPA)*, pages 160–170, Rye Brook, NY, July 2005.

[30] S. S. Chikkerur. Online fingerprint verification system. Master's thesis, State University of New York at Buffalo, 2005.

[31] C. K. Chow. An optimum character recognition system using decision functions. *IRE Transactions on Electronic Computers*, EC-6:247–254, December 1957.

[32] C. K. Chow. On optimum recognition error and rejection tradeoff. In *IEEE Transactions on Information Theory*, volume 16, pages 41–46, January 1970.

[33] D. Chralampidis, T. Kasparis, and M. Georgiopoulos. Classification of noisy signals using fuzzy ARTMAP neural networks. *IEEE Transactions on Neural Networks*, 12(5):1023 – 1036, 2001.

[34] L. P. Cordella, P. Foggia, C. Sansone, F. Tortorella, and M. Vento. Classification reliability and its use in multiclassifier systems. In *Lecture Notes in Computer Science*, volume 1310, pages 46–53. Springer, 1997.

[35] L. P. Cordella, P. Foggia, C. Sansone, F. Tortorella, and M. Vento. Optimizing the error/reject trade-off for a multi-expert system using the Bayesian combining rule. In A. Amin, D. Dori, P. Pudil, and H. Freeman, editors, *Lecture Notes in Computer Science*, volume 1451, pages 716–725, Berlin, 1998. Springer.

[36] J. Czyz, M. Sadeghi, J. Kittler, and L. Vandendorpe. *Biometric Authentication*, volume 3072 of *Lecture Notes in Computer Science*, chapter Decision Fusion for Face Authentication, pages 686–693. Springer, 2004.

[37] R. D'Agostino and M. Stephens. *Goodness-of-Fit Techniques*. Marcel Dekker, Inc., 1986.

[38] M. Dash and H. Liu. Feature selection for classification. *Intelligent Data Analysis*, 1:131–156, 1997.

[39] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B*, 39(1):1–38, 1977.

[40] Luc Devroye, Laszlo Gyorfi, and Gabor Lugosi. *A Probabilistic Theory of Pattern Recognition*, volume 25. Springer, 1996.

[41] G. Doddington, W. Liggett, A. Martin, M. Przybocki, and D. Reynolds. Sheeps, goats, lambs and wolves: A statistical analysis of speaker performance in the NIST 1998 Speaker Recognition Evaluation. In *Proceedings of the International Conference on Speech and Language Processing, ICSLP'98*, 1998.

[42] R. O. Duda, P. E. Hart, and D. G. Stork. *Pattern Classification*. Wiley Interscience, New York, 2nd edition, 2001.

[43] R. P. W. Duin. The combining classifier: to train or not to train? In *Proceedings of the International Conference on Pattern Recognition*, Quebec, Canada, 2002.

[44] M. El-Maliki. *Speaker Verification with Missing Features in Noisy Environments*. PhD thesis, Swiss Federal Institute of Technology Lausanne (EPFL), 2000.

[45] M. Falk, J. Hüsler, and R.-D. Reiss. *Laws of Small Numbers: Extremes and Rare Events*. Birkhäuser, 2nd edition, 2004.

[46] J. Fierrez, J. Ortega, J. Gonzalez, and J. Bigun. Kernel-based multimodal biometric verification using quality signals. *Biometric Technologies for Human Identification, Proceedings of SPIE*, 5404:544–554, 2004.

[47] J. Fierrez, J. Ortega-Garcia, D. Torre-Toledano, and J. Gonzalez-Rodriguez. BioSec baseline corpus: A multimodal biometric database. *Pattern Recognition*, 40(4):1389–1392, April 2007.

[48] J. Fierrez-Aguilar. *Adapted Fusion Schemes for Multimodal Biometric Authentication*. PhD thesis, Universidad Politecnica de Madrid, 2006.

[49] J. Fierrez-Aguilar, Y. Chen, J. Ortega-Garcia, and A. K. Jain. Incorporating image quality in multi-algorithm fingerprint verification. In *Proc. of the International Conference on Biometrics*, Hong Kong, January 2006.

[50] J. Fierrez-Aguilar, D. Garcia-Romero, J. Ortega-Garcia, and J. Gonzalez-Rodriguez. Adapted user-dependent multimodal biometric authentication exploiting general information. *Pattern Recognition Letters*, 26(16):2628–2639, 2005.

[51] J. Fierrez-Aguilar, J. Ortega-Garcia, J. Gonzalez-Rodriguez, and J. Bigun. Discriminative multimodal biometric authentication based on quality measures. *Pattern Recognition*, 38:777–779, 2005.

[52] G. Fumera and F. Roli. Analysis of error-reject trade-off in linearly combined multiple classifiers. *Pattern Recognition*, 37(6):1245–1265, 2004.

[53] G. Fumera and F. Roli. A theoretical and experimental analysis of linear combiners for multiple classifier systems. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(6):942–956, June 2005.

[54] G. Fumera, F. Roli, and G. Giacinto. Multiple reject thresholds for improving classification reliability. In *Proceedings of SSPR/SPR*, pages 863–871, 2000.

[55] G. Fumera, F. Roli, and G. Vernazza. Analysis of error-reject trade-off in linearly combined classifiers. In *Proc. 16th International Conference on Pattern Recognition*, volume 2, pages 120–123, Quebec, Canada, 2002.

[56] M. J. F. Gales and S. J. Young. Robust speech recognition in additive and convolutional noise using parallel model combination. *Computer Speech and Language*, 9:289–307, 1995.

[57] F. Galton. *Finger Prints*. MacMillan and Co., London, 1892.

[58] C.I Garris, C Watson, R. McCabe, and C.L Wilson. User's guide to NIST fingerprint image software. Technical Report NISTIR 6813, NIST, http://fingerprint.nist.gov/NFIS/index.html, November 2001.

[59] M. S. Gazzaniga, R. B. Ivry, and G. R. Mangun. *Cognitive Neuroscience, The Biology of the Mind*. Norton, 2nd edition, 2002.

[60] R. N. Giere. *Logic, Methodology and Philosophy of Science IV*, chapter Objective Single-Case Probabilities and the Foundations of Statistics. North-Holland, 1973.

[61] D. Gillies. Varieties of propensity. *British Journal for the Philosophy of Science*, 51:807–835, 2000.

[62] M. Golfarelli, D. Maio, and D. Maltoni. On the error-reject trade-off in biometric verification systems. In *IEEE Transactions on Pattern Analysis and Machine Intelligence*, volume 19, 1997.

[63] C. M. Grinstead and J. Laurie Snell. *Introduction to Probability*. American Mathematical Society, 1997.

[64] R. Gross and V. Brajovic. An image preprocessing algorithm for illumination invariant face recognition. In *Proceedings of the 4th International Conference on Audio- and Video-Based Biometric Person Authentication (AVBPA)*, Guilford, UK, 2003.

[65] P. Grother and E. Tabassi. Performance of biometric quality measures. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(4):531–543, April 2007.

[66] I. Guyon, C. Aliferis, and A. Elisseeff. Causal feature selection. In Huan Liu and Hiroshi Motoda, editors, *Computational Methods of Feature Selection*. Chapman and Hall/CRC Press, 2007.

[67] I. Guyon, S. Gunn, A. Ben Hur, and G. Dror. *Result Analysis of the NIPS 2003 Feature Selection Challenge*. MIT Press, Cambridge, MA, 2005.

[68] I. Guyon, S. Gunn, M. Nikravesh, and L. Zadeh, editors. *Feature Extraction, Foundations and Applications*. Studies in Fuzziness and Soft Computing. Springer Verlag, 2006.

[69] M. Haindl, J. Kittler, and F. Roli, editors. *Multiple classifier systems*. Number 4472 in LNCS. Springer, Prague, Czech Republic, 2007.

[70] D. J. Hand. Recent advances in error rate estimation. *Pattern Recognition Letters*, 4(5):335–346, 1986.

[71] A. Handjek. Interpretations of probability. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. 2007. http://plato.stanford.edu/.

[72] J. Hawthorne. Degree-of-belief and degree-of-support: Why Bayesians need both notions. *Mind*, 114(454):277–320, 2005.

[73] R.-L. V. Hsu, J. Shah, and B. Martin. Quality assessment of facial images. In *Proceedings of the 2006 Biometrics Symposium*, Baltimore, MD, USA, 2006.

[74] C. Huygens. *Van Rekeningh in Spelen van Geluck vertaald en toegelicht door Wim Kleijne*. Epsilon Uitgaven, Utrecht, 1660, 1998.

[75] E. C. Ifeachor and B. W. Jervis. *Digital Signal Processing, A Practical Approach*. Prentice Hall, 2002.

[76] A. K. Jain, Y. Chen, and M. Demirkus. Pores and ridges: High-resolution fingerprint matching using level 3 features. *IEEE Trans. Pattern Anal. Mach. Intell.*, 29(1):15–27, 2007. Anil K. Jain and Yi Chen and Meltem Demirkus.

[77] A. K. Jain, K. Nandakumar, and A. A. Ross. Score normalization in multimodal biometric systems. *Pattern Recognition*, 38(12):2270–2285, 2005.

[78] A. K. Jain and A. Ross. Multibiometric systems. *Commun. ACM*, 47(1):34–40, 2004.

[79] R. Jain, R. Kasturi, and B. G. Schunck. *Machine Vision*. McGraw Hill, 1995.

[80] X. Jiang, M. Liu, and A. C. Kot. Fingerprint retrieval for identification. *IEEE Transactions on Information Forensics and Security*, 1(4):532–542, December 2006.

[81] N. D. Kalka, J. Zuo, N. A. Schmid, and B. Cukic. Image quality assessment for iris biometric. In *Proceedings of SPIE, the International Society for Optical Engineering, Biometric technology for human identification*, volume 6202, pages D.1–D.11, 2006.

[82] J. F. Kenney and E. S. Keeping. *Mathematics of Statistics*. Van Nostrand, Princeton, NJ, 1951.

[83] J. Kittler, M. Hataf, R. Duin, and J. Matas. On combining classifiers. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(3):226–239, 1998.

[84] O. Koval, S. Voloshynovskiy, and T. Pun. Error exponent analysis of person identification based on fusion of dependent/independent modalities. In *In Proceedings of SPIE Photonics West, Electronic Imaging 2006, Multimedia Content Analysis, Management, and Retrieval 2006 (EI122)*, 2006.

[85] K. Kryszczuk and A Drygajlo. Addressing the vulnerabilities of likelihood-ratio-based face verification. In *Proc. 5th AVBPA*, Rye Brook NY, USA., 2005.

[86] K. Kryszczuk and A. Drygajlo. Gradient-based image segmentation for face recognition robust to directional illumination. In *Visual communications and image processing 2005 : 12-15 July 2005, Beijing, China*, 2005.

[87] K. Kryszczuk and A. Drygajlo. On combining evidence for reliability estimation in face verification. In *Proc. of the EUSIPCO 2006*, Florence, September 2006.

[88] K. Kryszczuk and A. Drygajlo. On face quality measures. In *Proc. 2nd Workshop on Multimodal User Authentication MMUA'06*, Toulouse, France, May 2006.

[89] K. Kryszczuk and A. Drygajlo. Improving classification with class-independent quality measures: Q-stack in face verification. In *Proc. of the 2nd International Conference on Biometric ICB'07*, Seoul, Korea, 2007.

[90] K. Kryszczuk and A. Drygajlo. Q-stack: uni- and multimodal classifier stacking with quality measures. In *Proceedings of the International Workshop on Multiple Classifier Systems*, Prague, Czech Republic, May 2007.

[91] K. Kryszczuk and A. Drygajlo. Reliability estimation for multimodal error prediction and fusion. In *Proc. of the 7th International Conference on Pattern Recogntition in Information Systems*, Funchal, Portugal, 2007.

[92] K. Kryszczuk, P. Morier, and A. Drygajlo. Study of the distinctiveness of level 2 and level 3 features in fragmentary fingerprint comparison. In *International Conference on Computer Vision, Biometric Authentication Workshop*, Prague, Czech Republic, May 2004.

[93] K. Kryszczuk, J. Richiardi, P. Prodanov, and A. Drygajlo. Error handling in multimodal biometric systems using reliability measures. In *13th European Signal Processing Conference (EUSIPCO 2005)*, Antalya, Turkey, 2005.

[94] K. Kryszczuk, J. Richiardi, P. Prodanov, and A. Drygajlo. Reliability-based decision fusion in multimodal biometric verification systems. *EURASIP Journal on Advances in Signal Processing*, 2007. Article ID 86572.

[95] L. I. Kuncheva. *Combining Pattern Classifiers: Methods and Algorithms*. John Wiley & Sons, 2004.

[96] L. I. Kuncheva and C. J. Whitaker. Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy. *Machine Learning*, 51(2):181–207, May 2003.

[97] P. Langley. Selection of relevant features in machine learning. In *Proc. Of AAAI Symposium on Relevance*, Technical Report FS-94-02, Menlo Park, CA, USA, 1994.

[98] J. Li and S. K. Zhou. Probabilistic face recognition from compressed imagery. In *Proceedings, ICASSP'04, IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 5, pages V – 909–12, May 2004.

[99] S. Z. Li and A. K. Jain, editors. *Handbook of Face Recognition*. Springer-Verlag, 2004.

[100] W. Li, G. Xiang, and B. Terrance. Predicting biometric system failure. In *CIHSPS 2005 - IEEE International Conference on Computational Intelligence for Homeland Security and Personal Safety*, Orlando, FL, March 2005.

[101] E. Lim, X. Jiang, and W. Yau. Fingerprint quality and validity analysis. In *Image Processing. 2002. Proceedings. 2002 International Conference on*, volume 1, pages 469–472, Rochester, NY, USA, 2002.

[102] E. Locard. Les pores et l'identification des criminels. *Biologica, Revue Scien-tifique de Medecin*, 22:357–362, 1912.

[103] D. Maio, D. Maltoni, R. Cappelli, J.L. Wayman, and A.K. Jain. FVC2002: Second fingerprint verification competition. In *Proc. of the 16th International Conference on Pattern Recognition*, Hong Kong, 2002.

[104] D. Maltoni and A. K. Jain, editors. *Biometric Authentication, Proceedings of the ECCV 2004 International Workshop BioAW 2004*. LNCS 3087. Springer, Prague, Czech Republic, 2004.

[105] D. Maltoni, D. Maio, A. K. Jain, and S. Prabhakar. *Handbook of Fingerprint Recognition*. Springer Professional Computing. Springer, 2003.

[106] A. M. Martinez. Recognizing imprecisely located, partially ocluded and expression variant faces from a single sample per class. *IEEE Transactions on Pattern Recognition and Machine Intelligence*, 24(6):748–763, 2002.

[107] K. Messer, J. Kittler, M. Sadeghi, M. Hamouz, A. Kostyn, S. Marcel, S. Bengio, F. Cardinaux, C. Sanderson, N. Poh, Y. Rodriguez, K. Kryszczuk, J. Czyz, L. Vandendorpe, J. Ng, H. Cheung, and B. Tang. Face authentication competition on the BANCA database. In *Proceedings of the ICBA*, Hong Kong, 2004.

[108] J. Meynet and J.-P. Thiran. Information theoretic combination of classifiers with application to AdaBoost. In M. Haindl, J. Kittler, and F. Roli, editors, *Multiple Classifier Systems, 7th International Workshop*, LNCS 4472, pages 171–179, Prague, Czech Republic, 2007.

[109] K. Michalak and H. Kwasnicka. Correlation-based feature selection strategy in classification problems. *International Journal of Applied Mathematics and Computer Scieince*, 16(4):503–511, 2006.

[110] M. Minami and K. Shimizu. Estimation of similarity measure for multivariate normal distributions. *Environmental and Ecological Statistics*, 6:229–248, 1999.

[111] T. M. Mitchell. *Machine Learning*. McGraw Hill, 1997.

[112] Karthik Nandakumar, Yi Chen, Sarat C. Dass, and Anil K. Jain. Biometric score fusion: Likelihood ratio, matcher correlation and image quality. Technical Report MSU-CSE-07-18, Michigan State University, March 2007.

[113] W. Peng and J. Qiang. Performance modeling and prediction of face recognition systems. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, New York, NY, USA, June 2006.

[114] P. J. Philips, P. Grother, R. J. Micheals, D. M. Blackburn, E. Tabassi, and M. Bone. Face recognition vendor test 2002, overview and summary. NIST, March 2003.

[115] P. J. Phillips. Support vector machines applied to face recognition. In S. A. Solla M. J. Kearns and D. A. Cohn, editors, *Advances in Neural Information Processing Systems*, volume 11. MIT Press, 1999.

[116] S. Pinker. *How The Mind Works*. W.W. Norton, 1997.

[117] N. Poh. *Multi-system Biometric Authentication: Optimal Fusion and User-Specific Information*. PhD thesis, Swiss Federal Institute of Technology Lausanne (EPFL), 2006.

[118] N. Poh and S. Bengio. Can chimeric persons be used in multimodal biometric authentication experiments? In *Proc. 2nd Joint AMI/PASCAL/IM2/M4 Workshop on Multimodal Interaction and Related Machine Learning Algorithms (MLMI)*, volume LNCS 3869, pages 87–100, Edinburgh, UK, 2005.

[119] N. Poh and S. Bengio. How do correlation and variance of base-experts affect fusion in biometric authentication tasks? *IEEE Transactions on Signal Processing*, 53(11):4384–4396, November 2005.

[120] N. Poh and S. Bengio. Improving fusion with margin-derived confidence in biometric authentication tasks. In *Proceedings of the AVBPA*, Rye Brook NY, USA, 2005.

[121] N. Poh, G. Heusch, and J. Kittler. On combination of face authentication experts by a mixture of quality dependent fusion classifiers. In *Proceedings of th 7th International Workshop on Multiple Classifier Systems*, Prague, Czech Republic, 2007.

[122] R. Polikar. Ensemble based systems in decision making. *IEEE Circuits and Systems Magazine*, 6(3):21–45, 2006.

[123] K. R. Popper. The propensity interpretation of probability. *British Journal for the Philosophy of Science*, 10:25–42, 1959.

[124] P. Prodanov and A. Drygajlo. Decision networks for repair strategies in speech-based interaction with mobile tour-guide robots. In *Proceedings of the 2005 IEEE International Conference on Robotics and Automation*, pages 3041– 3046, 2005.

[125] N. K. Ratha and R. Bolle. Fingerprint image quality estimation. In *Proc. of ACCV 2000*, pages 819–823, Taipei, Taiwan, 2000.

[126] D. A. Reynolds. *A Gaussian Mixture Modeling Approach to Text-Independent Speaker Identification*. PhD thesis, Georgia Institute of Technology, 1992.

[127] D. A. Reynolds. Channel robust speaker verification via feature mapping. In *Proceedings IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 03)*, volume 2, pages 53–56, Hong Kong, 2003.

[128] J. Richiardi, K. Kryszczuk, and A. Drygajlo. Quality measures in unimodal and multimodal biometric verification. In *Proceedings of the 15th European Conference on Signal Processing EUSIPCO 2007*, Poznan, Poland, September 2007.

[129] J. Richiardi, P. Prodanov, and A. Drygajlo. A probabilistic measure of modality reliability in speaker verification. In *Proc. IEEE International Conf. on Acoustics, Speech and Signal Processing 2005*, pages 709–712, Philadelphia, USA, March 2005. Winner of Best Student Paper Competition.

[130] J. Richiardi, P. Prodanov, and A. Drygajlo. Speaker verification with confidence and reliability measures. In *2006 IEEE International Conference on Acoustics, Speech and Signal Processing, 2006*, volume 1, pages 641–644, Toulouse, France, 2006.

[131] F. Roli, G. Fumera, and J. Kittler. Fixed and trained combiners for fusion of imbalanced pattern classifiers. In *Proc. of the Intl. Conf. on Information Fusion*, pages 278–284, Annapolis, MD, USA, 2002.

[132] F. Roli and J. Kittler, editors. *Multiple classifier systems*. Number 2364 in LNCS. Springer Verlag, Cagliari, Italy, 2002.

[133] A. Ross and A. K. Jain. Multimodal biometrics: An overview. In *Proc. of 12th European Signal Processing Conference (EUSIPCO)*, pages 1221–1224, Vienna, Austria, September 2004.

[134] A. A. Ross, K. Nandakumar, and A. K. Jain. *Handbook of Multibiometrics*. Springer, 2006.

[135] S. Russel and P. Norvig. *Artificial Intelligence. A Modern Approach*. Prentice Hall, 1995.

[136] M. Sadeghi, J. Kittler, A. Kostin, and K. Messer. A comparative study of automatic face verification algorithms on the BANCA database. In *Proc. of Audio- and Video-Based Biometric Person Authentication (AVBPA03)*, volume 2688/2003 of *Lecture Notes in Computer Science*, pages 35–43, Berlin, 2003. Springer.

[137] C. Sanderson. *Automatic Person Verification Using Speech and Face Information*. PhD thesis, Griffith University, Queensland, Australia, 2002.

[138] C. Sanderson and K. K. Paliwal. Likelihood normalization for face authentication in variable recording conditions. In *Proc. IEEE International Conference on Image Processing (ICIP)*, Rochester, NY, USA, 2002.

[139] C. Sanderson and K. K. Paliwal. Features for robust face based identity verification. *Signal Processing*, 83(5):931–940, 2003.

[140] M. Savvides and B.V.K. Vijaya Kumar. Illumination normalization using logarithm transforms for face authentication. In J. Kittler and M.S. Nixon, editors, *LNCS 2688: Proceedings of the 4th International Conference on Audio- and Video-Based Biometric Person Authentication (AVBPA)*, pages 549–556, Berlin, 2003. Springer-Verlag.

[141] R. Schapire. Strength of weak learnability. *Journal of Machine Learning*, 5:197–227, 1990.

[142] L. L. Shen, A. Kot, and W. M. Koo. Quality measures of fingerprint images. In *Proc. of the 3rd International Conference, AVBPA 2001*, volume 2091 of *LNCS*, pages 266–271, 2001.

[143] J. Short, J. Kittler, and K. Messer. Photometric normalisation for face verification. In *Lecture Notes in Computer Science, Audio- and Video-Based Biometric Person Authentication*, volume 3546/2005, pages 617–626. Springer-Verlag, Berlin, 2005.

[144] D. Simon-Zorita, J. Ortega-Garcia, J. Fierrez-Aguilar, and J. Gonzalez-Rodriguez. Image quality and position variability assessment in minutiae-based fingerprint verification. In *IEE Proceedings, Vision, Image and Signal Processing, Volume 150, Issue 6, 15 Dec. Page(s): 402 - 408*, volume 150, 2003.

[145] L. Sirovich and M. Kirby. Low-dimensional procedure for characterization of human faces. *Journal of the Optical Society of America*, 4(3):519–524, 1987.

[146] J. Q. Smith. *Decision Analysis: A Bayesian Approach*. Chapman and Hall, 1988.

[147] M. Sonka, V. Hlavac, and R. Boyle. *Image Processing, Analysis, and Machine Vision*. PWS Publishing, 2 edition, 1999.

[148] J. D. Stosz and L. A. Alyea. Automated system for fingerprint authentication using pores and ridge structure. In R. J. Mammone and J. D. Murley, editors, *Automatic Systems for the Identification and Inspection of Humans*, volume 2277, pages 210–223, San Diego, CA, 1994. SPIE.

[149] E. Tabassi and C. L. Wilson. A novel approach to fingerprint image quality. In *IEEE International Conference on Image Processing (ICIP)*, volume 2Genoa, Italy, September 2005., pages 37–40, Genoa, Italy, September 2005.

[150] E. Tabassi, C.L. Wilson, and C. Watson. NIST fingerprint image quality. Technical Report NISTIR 7151, NIST, August 2004.

[151] D. M. J. Tax and R. P. W. Duin. Using two-class classifiers for multiclass classification. In *Proceedings, 16th International Conference on Pattern Recognition*, volume 2, pages 124– 127, Quebec, Canada, 2002.

[152] S. Theodoridis and K. Koutroumbas. *Pattern Recognition, Second Edition*. Elsevier, 2003.

[153] K. A. Toh, W.-Y. Yau, E. Lim, L. Chen, and C.-H. Ng. Fusion of auxiliary information for multi-modal biometrics authentication. In *Proceedings of International Conference on Biometrics*, Lecture Notes in Computer Science, pages 678–685, Hong Kong, 2004. Springer.

[154] G. T. Toussaint. Bibliography on estimation of misclassification. *IEEE Transactions on Information Theory*, 20:472–479, 1974.

[155] K. Tumer and J. Ghosh. Theoretical foundations of linear and order statistics combiners for neural pattern classifiers. Technical report, Technical Report TR-95-02-98, Computer and Vision Research Center, University of Texas, Austin, 1995.

[156] K. Tumer and J. Ghosh. *Combining Artificial Neural Nets: Ensemble and Modular Multi-Net Systems*, chapter Linear and order statistics combiners for pattern classification, pages 127–162. Springer-Verlag, 1999.

[157] M. A. Turk and A. P. Pentland. Eigenfaces for recognition. *Journal of Cognitive Neuroscience*, 3(1):71–86, 1991.

[158] M. A. Turk and A. P. Pentland. Face recognition using eigenfaces. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 586–591, Maui, Hawaii, 1991.

[159] J. Venn. *The Logic of Chance*. Macmillan and Co., New York, 2 edition, 1876. reprinted 1962.

[160] R. Wang and B. Bhanu. Learning models for predicting recognition performance. In *Proceedings of the Tenth IEEE International Conference on Computer Vision (ICCV)*, volume 2, pages 1613–1618, Beijing, China, 2005.

[161] R. Wang, B. Bhanu, and H. Chen. An integrated prediction model for biometrics. In *Proceedings International Conference on Audio- and Video-based Biometric Person Authentication*, Rye Brook, NY, July 2005.

[162] C. I. Watson, M. D. Garris, E. Tabassi, C. L. Wilson, R. M. McCabe, and S. Janet. User's guide to NIST fingerprint image software 2. Technical Report NISTIR 6813, NIST, http://fingerprint.nist.gov/NFIS/index.html, November 2004.

[163] L. Wein and M. Baveja. Using fingerprint image quality to improve the identification performance of the U.S. VISIT program. In *Proceedings of the National Academy of Sciences, 2005*, 2005.

[164] I. H. Witten, A. Moffat, and T. C. Bell. *Managing Gigabytes - Compressing and Indexing Documents and Images*. Morgan Kaufmann, 1999.

[165] D. Wolpert. Stacked generalization. *Neural Networks*, 5:241–259, 1992.

[166] D. Wu, J. Koreman, and A. C. Morris. Discriminative features by MLP preprocessing for robust speaker recognition in noise. In *Proceedings of the ESSV (Elektronische Sprachsignalverarbeitung)*, Wolfenbuttel, Germany, 2005.

[167] S. L. Zabell. Fingerprint evidence. *Journal of Law and Policy*, 13(1):143–179, 2005.

[168] W. Zhao, R. Chellappa, A. Rosenfeld, and P. Phillips. Face recognition: A literature survey. *ACM Computing Surveys (CSUR)*, 35:399–458, 2003.

[169] S. Zhou and R. Chellappa. From sample similarity to ensemble similarity: Probabilistic distance measures in reproducing kernel Hilbert space. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(6):917–929, 2006.

Curriculum Vitae

Krzysztof Kryszczuk

Education

- 2003 - 2007: **Doctorate in Pattern Recognition**
  Swiss Federal Institute of Technology (EPFL) , Lausanne, Switzerland
- 1999 - 2001: **Master of Science in Psychology (Cognitive Science Engineering)**
  Rensselaer Polytechnic Institute, Troy, NY, USA
- 1993 - 1999: **Bachelor/Master of Science in Electrical Engineering**
  Lublin Technical University, Lublin, Poland

Awards

- 2007: Winner of the 2007 European Biometrics Forum Research Award
- 2005: Finalist of the EUSIPCO 2005 Best Student Paper competition
- 1999: Lutron Electronics Scholarship
- 1995: Dean's Award for academic merit, Department of EE, Lublin Technical University

Professional experience

- 2003 - 2007: **Engineer/Research Assistant**
  Swiss Federal Institute of Technology (EPFL), Lausanne, Switzerland
- 2002 - 2003: **Research Assistant**
  University of Geneva, Geneva, Switzerland
- 2001 - 2002: **Research Engineer**
  National University of Singapore, Singapore
- summer 2000: **Intern**
  Motorola, Inc., Motorola Research Labs, Plantation, FL, USA
- 1999 - 2001: **Research/Teaching Assistant**
  Rensselaer Polytechnic Institute (RPI), Troy, NY, USA

- 1998: **Consultant, Lighting Systems**
  Golland Sp.z.o.o. representative of Sylvania Lighting International (SLI) , Gorzow, Poland

- 1997,1998: **2 Internships, Lighting Systems**
  6 months in total, Sylvania Lighting International (SLI), Erlangen, Germany

Publications

- J. Kittler, N. Poh, O. Fatukasi, K. Messera, K. Kryszczuk, J. Richiardi, and A. Drygajlo. Quality dependent fusion of intramodal and multimodal biometric experts. In *Proceedings of SPIE 2007*, 2007.

- K. Kryszczuk and P. R. Boyce. Detection of slow light level reduction. *Journal of the IESNA*, 31(2):3–10, 2002.

- K. Kryszczuk and A. Drygajlo. Color correction for face detection based on human visual perception metaphor. In *Proceedings of the Workshop on Multimodal User Authentication, MMUA*, pages 138–143, Santa Barbara, CA, USA, 2003.

- K. Kryszczuk and A Drygajlo. Addressing the vulnerabilities of likelihood-ratio-based face verification. In *Proc. 5th AVBPA*, Rye Brook NY, USA., 2005.

- K. Kryszczuk and A. Drygajlo. Robust method of reference point localization in fingerprints. In *Proceedings of the COST 275 Workshop on Biometrics on the Internet*, Hatfield, UK, October 2005.

- K. Kryszczuk and A. Drygajlo. Gradient-based image segmentation for face recognition robust to directional illumination. In *Visual communications and image processing 2005 : 12-15 July 2005, Beijing, China*, 2005.

- K. Kryszczuk and A. Drygajlo. On combining evidence for reliability estimation in face verification. In *Proc. of the EUSIPCO 2006*, Florence, September 2006.

- K. Kryszczuk and A. Drygajlo. Singular point detection in fingerprints using quadrant change information. In *Proceedings of the International Conference on Pattern Recognition (ICPR)*, Hong Kong, China., August 2006.

- K. Kryszczuk and A. Drygajlo. On face quality measures. In *Proc. 2nd Workshop on Multimodal User Authentication MMUA'06*, Toulouse, France, May 2006.

- K. Kryszczuk and A. Drygajlo. Improving classification with class-independent quality measures: Q-stack in face verification. In *Proc. of the 2nd International Conference on Biometric ICB'07*, Seoul, Korea, 2007.

- K. Kryszczuk and A. Drygajlo. Q-stack: uni- and multimodal classifier stacking with quality measures. In *Proceedings of the International Workshop on Multiple Classifier Systems*, Prague, Czech Republic, May 2007.

- K. Kryszczuk and A. Drygajlo. Reliability estimation for multimodal error prediction and fusion. In *Proc. of the 7th International Conference on Pattern Recogntition in Information Systems*, Funchal, Portugal, 2007.

- K. Kryszczuk, A. Drygajlo, and P. Morier. Extraction of level 2 and level 3 features for fragmentary fingerprint comparison. In *Proceedings of the COST275*, Vigo, Spain, March 2004.
- K. Kryszczuk, P. Morier, and A. Drygajlo. Study of the distinctiveness of level 2 and level 3 features in fragmentary fingerprint comparison. In *International Conference on Computer Vision, Biometric Authentication Workshop*, May 2004.
- K. Kryszczuk, J. Richiardi, P. Prodanov, and A. Drygajlo. Error handling in multimodal biometric systems using reliability measures. In *13th European Signal Processing Conference (EUSIPCO 2005)*, Antalya, Turkey, 2005.
- K. Kryszczuk, J. Richiardi, P. Prodanov, and A. Drygajlo. Reliability-based decision fusion in multimodal biometric verification systems. *EURASIP Journal on Advances in Signal Processing*, 2007. Article ID 86572.
- K. Messer, J. Kittler, M. Sadeghi, M. Hamouz, A. Kostyn, S. Marcel, S. Bengio, F. Cardinaux, C. Sanderson, N. Poh, Y. Rodriguez, K. Kryszczuk, J. Czyz, L. Vandendorpe, J. Ng, H. Cheung, and B. Tang. Face authentication competition on the BANCA database. In *Proceedings of the ICBA*, Hong Kong, 2004.
- M. Perkins, B. Stengel, B. O'Dea, and K. Kryszczuk. In-building location systems are realized with fluorescent lamp infrastructures. *Journal of the IES*, 31(1):61–69, 2002.
- J. Richiardi, K. Kryszczuk, and A. Drygajlo. Quality measures in unimodal and multimodal biometric verification. In *Proceedings of the 15th European Conference on Signal Processing EUSIPCO 2007*, Poznan, Poland, September 2007.
- K. Kryszczuk and A. Drygajlo. Credence Measures and Error Prediction in Biometric Identity Verification. In *Signal Processing*, http://dx.doi.org/10.1016/j.sigpro.2007.10.007